University of Alberta

Three Approaches to Investigating the Multidimensional Nature of a Science Assessment

by

Rebecca Jayne Gokiert ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta
Fall, 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-32961-0
Our file Notre référence
ISBN: 978-0-494-32961-0

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

The purpose of this study was to investigate a multi-method approach for collecting validity evidence about the underlying knowledge and skills measured by a large-scale science assessment. The three approaches included analysis of dimensionality, differential item functioning (DIF), and think-aloud interviews. The specific research questions addressed were: (1) Does the 4-factor model previously found by Hamilton et al. (1995) for the grade 8 sample explain the data? (2) Do the performances of male and female students systematically differ? Are these performance differences captured in the dimensions? (3) Can think-aloud reports aid in the generation of hypotheses about the underlying knowledge and skills that are measured by this test?

A confirmatory factor analysis of the 4-factor model revealed good model data fit for both the AB and AC tests. Twenty-four of the 83 AB test items and 16 of the 77 AC test items displayed significant DIF, however, items were found, on average, to favour both males and females equally. There were some systematic differences found across the 4-factors; items favouring males tended to be related to earth and space sciences, stereotypical male related activities, and numerical operations. Conversely, females were found to outperform males on items that required careful reading and attention to detail.

Concurrent and retrospective verbal reports (Ericsson & Simon, 1993) were collected from 16 grade 8 students (9 male and 7 female) while they solved 12 DIF items. Four general cognitive processing themes were identified from the student protocols that could be used to explain male and female problem solving. The themes included comprehension (verbal and visual), visualization, background knowledge/experience (school or life), and strategy use. There were systematic differences in cognitive

processing between the students that answered the items correctly and the students who answered the items incorrectly; however, this did not always correspond with the statistical gender DIF results. Although the multifaceted approach produced interpretable and meaningful validity evidence about the knowledge and skills, these forms of validity evidence only begin to provide a basic understanding of the underlying construct(s) that are being measured.

Dedication

I would like to dedicate this dissertation to my parents, MaryLynne and Stephen Gokiert, for your endless support and guidance throughout my life. You have taught me what to cherish in life and I hope that I can continue to make you proud in the future. I love you guys!

Acknowledgements

I would like to express my deepest appreciation to a few people whose involvement and friendship was especially valuable and made this process more enjoyable.

I would like to thank Dr. Jacqueline Leighton, my doctoral supervisor, for her advice, support, feedback, and encouragement throughout my PhD. My research interests were fostered in a large part by the opportunities I had working as a research assistant on a number of her projects. Her knowledge and insights have helped me to sharpen my thinking and raised my confidence in my research abilities. She has provided me with some amazing opportunities to publish and travel and I am truly grateful that I was able to have her as a supervisor.

Special thanks to Dr. Todd Rogers who took the time to help me realize that I was meant to study in CRAME and made the transition smooth. While his courses were scary at times, with all of the mathematical proofs, I truly appreciate his excitement for teaching and his willingness to spend countless hours of his time devoted to his students.

I would like to thank Dr. Mark Gierl who has been an integral part of my graduate career, as a member of both my Master's and PhD committees. His expertise and enthusiasm for teaching and research have been very contagious and have contributed to my excitement about research. I am grateful for the opportunities his has afforded me and specifically an all expense paid trip to New York - what more could a graduate student ask for.

Special thanks are due to Dr. Stephen Norris and Dr. Judy Lupart for taking the time to read my document thoroughly and for their insightful questions and understandings of my research. They both made the candidacy and final defense a little less scary – thank you. I appreciate the thoughtful comments and recommendations that were provided by Dr. Marielle Simon, my external committee member.

I am extremely appreciative of the support I received from Pierre Brochu; his quick responses when I had questions about the SAIP and his familiar face at my presentations at AERA and NCME. This study would not have been possible without the support of the administrators and students welcoming me into their school and making data collection fun.

Sincere thanks to all of the wonderful friends in my life and specifically at school. To the girls, I am extremely grateful for their support, the luncheons, dinners, celebrations, and opportunities to vent. I am happy that I also got to share in their life experiences such as weddings, births, and house purchases. Special thanks to Dr. Janine Odishaw, my sister-in-law and best friend, this experience would not have been the same without her by my side the whole way – I am truly shocked that we made it and even defended the same week (talk about timing). I have made some great friends and colleagues for life and for that I am truly thankful.

To my family, who have been patient but always asking me when they were going to get to call me Doctor, I am truly appreciative. Growing up in a family of 4 brothers contributed to my interest in gender differences, and I would like to thank each one of them for their contribution to my success. They have always supported me in everything that I do, providing a welcome diversion with family dinners, weekends at the lake, and skiing in the mountains. Now I will truly get to relax and enjoy myself with nothing hanging over my head. Though she can't appreciate it now, to my niece Abi, whose smiles and ability to say "aunty becca" never let me forget what's important.

To my husband Dustin, who joined me at the beginning of what we both thought would be a never ending journey; who ever said you can't write a dissertation and a thesis at the same time, renovate a house, and plan a wedding single-handedly – was right! However, if you do it together you can accomplish great things! Dustin's support through this process by dealing with any and all computer meltdowns, my meltdowns, and putting up with not having a workable dining room table until now, will always be appreciated – I love you!

Chapter One: Introduction **1**

Purpose of the Study 4
Organization of the Dissertation 5

Chapter Two: Review of the Literature **7**

Large-Scale Assessment and Accountability: An Overview 7
  Large-Scale Assessment in Science 9
Methods for Investigating Construct Validity 12
  Dimensionality and Science Assessment 14
  Differential Item Functioning and Science Assessment 20
    Gender differences in science assessment. 23
    Item format differences. 26
  Verbal Reports and Science 28
    Identifying knowledge and skills. 30
    Identifying group differences. 33
Summary of Literature Reviewed 35

Chapter Three: Methods and Results: Psychometric Approaches **38**

Data 38
  SAIP Science Assessment 38
  Sample 40
Confirmatory Factor Analysis 41
  The 4-factor Model (ES, SR, CK, and RK) 41
    AB and AC results. 42
Differential Item Functioning (DIF) 44
  DIF Method 44
  DIF Results 45
    AB results. 46
    AC results. 46
    DIF and factor structure. 46
    Selection of DIF items for think-aloud 48

Chapter Four: Method and Results: Verbal Reports **52**

Think-Aloud Verbal Reports 52
  Participants 52
  Interview 53
  Data Analysis 54
  Results 57
  Summary of Verbal Report Data 76

Chapter Five: Discussion and Conclusions **82**

Purpose and Research Questions 82
Method and Summary of Findings 83
Limitations 91
Conclusions 92
Implications for Practice and Future Research 93

References 95

Appendix A: Grade 8 Factor Descriptions 109

Appendix B: Links Between Items and Assessment Blueprint 111

Appendix C: 4-Factor Coding for the AB and AC Tests 112

Appendix D: DIF, Factor, and Item Format for the AB and AC Tests 114

Appendix E: Informed Consent 118

Appendix F: Confidentiality Agreement 120

Appendix G: Think-Aloud Instructions 121

Appendix H: Coding Scheme 122

Appendix I: Think-Aloud Protocols Across the 12 DIF Items 123

List of Tables

Table 1. Sample of grade 8 students that wrote the 2004 SAIP Science Assessment...... 41

Table 2. CFA of Hamilton et al.'s (1995) 4-factor Model Applied to SAIP Tests AB and AC... 44

Table 3. SIBTEST results for the AB and AC Tests... 46

Table 4. AB DIF items by gender, factor loading, and format................................ 47

Table 5. AC DIF items by gender, factor loading, and format................................ 48

Table 6. Items selected for protocol analysis.. 50

Chapter One: Introduction

Large-scale assessment[1] has become a national and international method for monitoring student achievement and for ensuring that educational systems are working (Alberta Education, 2005; Hamilton, Stecher, & Klein, 2002; McGehee & Griffith, 2001). Large-scale assessments such as those conducted as part of the Pan-Canadian Assessment Program (PCAP) formally known as the School Achievement Indicators Program (SAIP) and the National Assessment of Educational Progress (NAEP) in the United States, are used to assess student achievement across a range of subject areas. Traditionally, Canada has used large-scale assessments primarily to track students' progress over their school career and to determine whether a student should graduate from high school (Alberta Education, 2005). In the United States, however, with the introduction of new legislation, such as the No Child Left Behind Act, the US education system has started to use large-scale assessments predominantly to track student progress for accountability purposes (Chudowsky & Pellegrino, 2003; Lane, 2004; NCLB Act, 2001 – Public Law 107-110; Popham, 1999).

As society and governments place more emphasis on large-scale testing, it has become increasingly important to examine the quality of large-scale testing programs and, in particular, the validity of inferences drawn from large-scale testing. Validity, as defined in the *Standards for Educational and Psychological Testing* (AERA, APA, &, NCME, 1999, p. 9), is "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests." Recognizing the importance of test validation, Haladyna and Downing (2004) argued that defining the construct is the first step to be taken in order for appropriate test score interpretation to occur. Suitable test score interpretation rests on grounded *construct formulation* (Cronbach & Meehl, 1955), which comprises an explicit knowledge and understanding of the latent trait(s) or knowledge and skills to be measured by the test. Although there is some debate regarding how construct validation should be conceptualized and applied, there is agreement that evidential support should be gathered to enhance the validity of inferences that are drawn from test scores (Cronbach, 1988; Kane, 1992; Maguire, Hattie, & Haig, 1994; Messick, 1989). The proposed interpretation of a test score must also be defensible against alternative interpretations through the use of empirical and other forms of reasonable evidence (Shavelson & Ruiz-Primo, 2000). Ensuring the validity of inferences drawn from large-scale assessments is especially important given that the inferences drawn from these tests have multiple purposes including accountability, informing instruction, determining achievement at the individual, classroom, school, and state/provincial levels, and for making high-stakes decisions. These high-stakes decisions can include whether an individual will graduate from high school, be admitted into a professional program (e.g., graduate school, law school), or receive a professional designation (~~e.g.,~~ becoming a registered psychologist). When students perform poorly, even seemingly low-stakes large-scale assessments have potential consequences (Haladyna & Downing, 2004). It is especially important to examine the construct validity of a test when inferences are made about a particular student's academic strengths and weaknesses (National Research Council [NRC], 2001).

---

[1] For the purpose of this paper the terms large-scale assessment and large-scale test will be used interchangeably.

Construct validation is the process through which the suitability test score interpretations are examined in the context of the latent trait(s) or knowledge and skills measured by the test. It is often the case that developers of large-scale assessment tools fail to verify the types of skills that are measured by tests and fail to provide guidelines for how strengths and weaknesses in student performance should be interpreted (Messick, 1994; NRC, 2001). When developers of large-scale assessments fail to explicitly state and describe the skills that are measured by these tests, test score interpretation becomes problematic. This situation of ill-defined constructs could be improved by engaging in empirical studies that investigate the underlying knowledge and skills measured by tests (NRC, 2001). In order to investigate the underlying knowledge and skills measured by large-scale tests of achievement, different forms of validity evidence must be explored.

Although there are vast forms of evidence, the construct validity literature related to large-scale assessments has focused on dimensionality, differential item functioning (DIF), and think-aloud interviews. Much research has demonstrated that dimensionality can enhance the validity of the inferences made from test scores by providing a better understanding of the latent dimensional structure (the knowledge and skills) of the test (Ayala, Shavelson, Yin, & Shultz, 2002; Childs & Oppler, 2000; Frenette & Bertrand, 2000; Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Nussbaum, Hamilton, & Snow, 1997). When attempting to describe the dimensional structure of a test it is also important to identify and understand implicitly the dimensional structure of individual items. When items are found to measure multiple dimensions, this can result in differential item functioning (DIF) for groups of students. One of the most carefully examined forms of DIF in large-scale assessments is the performance differences between males and females (Ryan & DeMark, 2002). Finally, not only can small scale interviews shed light on the gender differences associated with test item performance

(Hamilton, 1999; Ercikan et al., 2004) they can also be used to examine the underlying cognitive skills that students employ in problem solving, lending credibility to both DIF and dimensionality results (Hamilton, 1999; Hamilton, Nussbaum, & Snow, 1997; Norris, 1990; Norris, Leighton, & Phillips, 2004).

Purpose of the Study

Science achievement, as measured by large-scale tests, is often characterized by a number of reasoning skills, for example, quantitative reasoning, scientific reasoning, and spatial-mechanical reasoning (e.g., Hamilton et al., 1995; Nussbaum et al., 1997). If these dimensions represent distinct knowledge and skills in scientific achievement, it is important that tests capture and test scores reflect an individual's performance across these domains. If the subject domain to be assessed by the tests is in fact measured it will yield student scores that can be validly interpreted in terms of the student's strengths and weaknesses in the different areas of science achievement.

The purpose of this study, therefore, was to use three approaches to collect evidence about the underlying knowledge and skills measured by the 2004 version of the School Achievement Indicators Program (SAIP) Science Assessment (Council of Ministers of Education, Canada [CMEC], 2000) now called the Pan-Canadian Assessment Program (PCAP). The SAIP is a low-stakes large-scale achievement assessment; therefore, it provides researchers the opportunity to examine the efficacy of collecting validity evidence within a low-risk environment. The three approaches for collecting validity evidence in the present study included the analysis of dimensionality, differential item functioning (DIF), and think-aloud interviews. The specific research questions addressed were:

(1) Does the 4-factor model previously found by Hamilton et al. (1995) for the grade 8 sample explain the SAIP 2004 data? If not, what is the dimensional structure of the 2004 version of the SAIP Science assessment?

(2) Do the performances of male and female students on SAIP items systematically differ? Are these performance differences systematically captured in the dimensions?

(3) If performance differences are found, can think-aloud reports of male and female grade 8 students aid in the generation of hypotheses about the underlying knowledge and skills that are measured by this test?

It is hypothesized that determining the underlying knowledge and skills that are measured by the SAIP Science Assessment with different statistical and substantive methods can inform and enhance the validity of inferences drawn about student performance. More generally, these methods can aid in elucidating the underlying reasoning skills that are measured by large-scale science tests.

Organization of the Dissertation

In Chapter 2, to familiarize the reader with large-scale assessments a brief background of large-scale assessments and accountability is provided. This is followed by a specific focus on national and international large-scale science assessments in grade 3 through grade 12. A discussion of how distinct methods, such as dimensionality analysis and DIF, can inform investigations of the construct validity of a test followed by a discussion of gender differences in science assessments and how using think-aloud reports for uncovering the source of these differences can inform construct validity are then presented. Finally, dimensionality, DIF, and think-aloud interviews as forms of validity evidence are discussed. In Chapter 3, the methods and results for the psychometric approaches are discussed in detail. This chapter includes a thorough description of the SAIP Science Assessment test and sample that were used for both the confirmatory factor analysis (CFA) and the differential item functioning (DIF) analyses. This is followed by a description of the CFA and DIF methods used and the results obtained in the study and how they are interrelated. Chapter 4 includes a detailed description of the think-aloud methods, sample of students involved in the interview portion of the study, and the think-aloud results. A summary of the purpose of the study, the research questions, and the results are provided in chapter 5 followed by a discussion of the conclusions and implications for practice and recommendations for future research.

Chapter Two: Review of the Literature

Large-Scale Assessment and Accountability: An Overview

In Canada educational accountability is operationalized through the use of large-scale provincial and national assessments to inform instruction, provide a method of comparing performance from province to province and school to school, and to provide test scores to teachers, parents, and students (Alberta Education, 2005; CMEC, 2000). Educational accountability in the United States is slightly different in that it refers to the use of results from large-scale assessments to inform instructional practices, reward teachers and schools for performance, provide test scores to teachers, parents and students, and to determine whether schools need to enter mandatory school-improvement programs (Cizek, 2001; Hamilton et al., 2002). There can be inherent risks and rewards when an educational accountability system is employed. For example, in the United States, if a school performs poorly on state mandated tests, teachers' wages may be affected (they may not receive bonuses) (Haladyna & Downing, 2004), student enrolment may drop as parents are given the option of withdrawing their children from a low performing school, and school morale may suffer. Alternatively, teachers may receive bonuses for good student performance and school districts may receive extra funding for school-based programs (Hamilton et. al., 2002).

For several decades, large-scale assessment has been widely used in most educational systems across North America (Hamilton et al., 2002). Large-scale assessments are typically developed and mandated by groups (e.g., government agencies) that are external to the schools and classrooms in which they are administered (Hamilton et al., 2002). Large-scale assessments can be administered to samples of students or all students, across districts, schools, and classrooms. Aside from accountability, these assessments are used for a number of purposes: to determine the level of student achievement; for college, graduate, and law or medical school admissions; and for the certification of professionals. Moreover, large-scale assessments can be high-stakes or low-stakes. Implicit in high-stakes assessment programs is the provision of rewards or sanctions such as the examples mentioned previously (Downing & Haladyna, 1996; Kane, 2002; Moss, 1998). Conversely, low-stakes tests are typically used to provide information regarding student performance, with little feedback to students, and few, if any, consequences associated with examinee scores (DeMars, 2000; Hamilton, 1998).

Large-scale achievement assessment is most common from grade 3 to grade 12. Across Canada, there are different standards for assessing student achievement throughout a student's school career. For example, in the provinces of Alberta and British Columbia, student achievement is assessed through the use of large-scale assessments at least four to five times from grade 3 to grade 12 (Alberta Education, 2005). More specifically, students are administered large-scale provincial achievement tests (PAT) in grades 3, 6, and 9, and diploma exams in grade 12 in the province of Alberta. High school graduates, pursuing postsecondary degrees in other countries, may be expected to write large-scale college admission tests such as the Scholastic Assessment Test [SAT] (College Board, 2005) in the USA. Although it is important to discuss the implications of large-scale assessments generally, the focus of this research will be on pursuing three avenues of study – dimensionality analysis, differential item functioning, and think-aloud interviews – to uncover the knowledge and skills measured by a large-scale science assessment in Canada.

Large-Scale Assessment in Science

A number of large-scale assessment tools have been developed to assess the science performance of students at local, national, and international levels. To understand the complex nature of student achievement in science, researchers frequently study student performance on national and international large-scale assessments. To follow, some examples of national and international large-scale assessments that are frequently studied will be presented. The National Education Longitudinal Study of 1988 (NELS: 88) includes achievement tests in science, mathematics, reading, and history and is used as a measure of student achievement. The science portion of the NELS: 88 test was initially administered to a national sample of 8th grade students and again to those same students when they were in grades 10 and 12 (Hamilton et al., 1997). Another large-scale assessment in the United States is the National Assessment of Educational Progress (NAEP), mandated by the United States Congress. The NAEP was initially developed as a low-stakes assessment program used to track achievement in the majority of subject areas over time; however, NAEP is now being used primarily as a research instrument to influence and shape state policy decisions (Johnson, 1999).

In Canada, the School Achievement Indicators Program (SAIP) is also a low-stakes test that is administered to probability samples of 13- and 16-year-old students in each province to allow reporting at the provincial level in a number of subject areas (science, writing, and mathematics). The SAIP was developed, administered, and reported by the Council of Ministers of Education, Canada (CMEC). CMEC is the national voice for education in Canada and represents nationally and internationally the educational welfare of the provinces and territories. CMEC used the SAIP Science

Assessment as a report card of Canadian students' knowledge and problem solving in science. The SAIP has been redefined and renamed as the Pan-Canadian Assessment Program (PCAP). The first PCAP assessment administration was to a random sample of 13-year-old students in schools across Canada in the spring of 2007, with reading as the major component, and science and math as the minor components (CMEC, 2007). The PCAP will measure these domains on a cyclical basis (every two to three years) for both 13- and 15-year-olds. Given that there is no school or student level reporting, the SAIP was a low-stakes assessment program. The science assessment portion of the SAIP was administered on three separate occasions (1996, 1999, 2004), and the results were used as a report card of students' scientific knowledge and reasoning skills at the provincial level across Canada.

The International Association for the Evaluation of Educational Achievement (IEA) administered the Trends in International Mathematics and Science Study (TIMSS) to measure trends in students' math and science achievement. Countries that participated in TIMSS have the opportunity to measure students' progress in mathematics and science achievement on a 4-year cycle (e.g., 1995, 1999, 2003, and spring 2007) (Johnson, 1999; National Centre for Education Statistics [NCES], 2007). The most recent administration of the TIMSS measured mathematics and science achievement at two levels (elementary and middle school) with participation from 46 countries in 2003 and an estimated 63 countries in 2007 (NCES, 2007). The TIMSS data have been used to evaluate the effectiveness of the participating countries' educational system and, in Canada, some of the provinces' educational systems.

The Organisation for Economic Co-operation and Development (OECD) administers the Programme for International Student Assessment (PISA), an internationally standardized assessment tool, to 15-year-old students across approximately 40 countries. The PISA was administered in 2000, 2003, and 2006 and assesses students' knowledge in reading, mathematics, and science problem solving. CMEC intends to harmonize the Pan-Canadian Assessment Program (PCAP) with the PISA administration schedule.

Beyond the use of national and international tests, provincial and state governing bodies mandate assessment of science typically at the elementary, middle, and high school levels. Each province and state has independent standards for science test development and implementation. For example, in Canada, test design and implementation are mandated at the provincial level (Alberta Education, 2005). Provincial tests are administered yearly. However, when evaluating a country's science achievement it seems logical that researchers and educational officials would be inclined to use national and international tests instead of state and provincial tests. The reasoning behind this is that national and international testing programs provide a standardized method for comparing achievement across different states, provinces, and countries.

Although a number of national and international tests are considered by many to be the gold standard in achievement testing (e.g., NAEP, PISA), these tests, which primarily focus on content domains such as reading, mathematics and science (Linn, 2002), may ignore the measurement of cognitive processing and reasoning skills within these content domains (Mosher, 2004). The constructs that are measured by large-scale science achievement tests are not well articulated or consistent across different science tests. For example, Mosher (2004) indicated that the NAEP science assessment measures some combination of ability and achievement that has not yet been closely examined or sorted out. Understanding what test scores represent is difficult when it is not clear what large-scale science achievement tests are measuring outside of the specific content described in test blueprints. Armed with test blueprints, researchers and educational stakeholders that are interested in examining the usefulness of test scores are required to provide the evidential support necessary to make inferences about student performance in science. To strengthen a testing program, it would seem logical that test developers take a leading role in integrating validity evidence into their testing programs, which will ultimately enhance the tests and the inferences that are generated from the test scores. Engaging in this rigorous process would enable the users of these tests to make direct statements about how well a student performed on a given construct and further the development of sound arguments regarding the validity of interpretation and use of test scores (Haladyna, 2002a).

Methods for Investigating Construct Validity

The view that construct validity is the most important standard against which tests and test scores should be evaluated has increasingly prompted test developers and researchers to examine this form of validity more closely (Cronbach, 1988; Kane, 1992; Messick, 1989; Shepard, 1993). Cronbach and Meehl (1955), in their seminal article entitled *Construct Validity in Psychological Tests,* described the concept of construct validation as a process which reflects “a particular construct, to which are attached certain meanings” (p. 65). They further suggested that testable hypotheses can be developed, based on the inferences drawn from test scores, which can be confirmed or disconfirmed. In its early stages, construct validity was considered "something for theoreticians" (Haladyna, 2002b, p. 485), and the practical application of construct validity to the interpretation of test scores was unclear. To address this, in the third edition of *Educational Measurement*, Messick (1989) provided practical ways for researchers and test developers to both conceptualize and approach test validation. Messick offered six aspects of validity evidence (content, structural, generalizability, external, consequential, and substantive), which would serve as the basis for collecting different types of evidence and for developing an overall validity argument. Development of the validity argument can be considered a process or approach whereby empirical and theoretical evidence is collected and applied to support the "appropriateness of inferences and actions based on test scores" (Linn, 2002. p. 31). Kane (2006), in the most recent edition of *Educational Measurement*, proposed that the validity argument provides an overall evaluation of the intended interpretations and uses of tests scores. He distinguished between validation as a process of evaluating whether a proposed interpretation or use of a test score is plausible, and validity as the extent to which evidence supports or contests the interpretations or uses. According to Haladyna and Downing (2004), the central issue in evaluating the appropriateness of inferences made from test scores is construct validity (see also Ryan & DeMark, 2002). Haladyna and Downing (2004) recommended the following three steps to generate a validity argument: "(a) create a plausible argument regarding a desired interpretation or use of a test score, (b) collect and organize validity evidence bearing on this argument, and (c) evaluate the argument and the evidence concerning the validity of the interpretation" (p. 19). Dimensionality analysis, DIF and think-aloud interviews are three forms of evidence that can be used in the formulation of the validity argument. In the sections to follow, dimensionality, DIF, and think-aloud interviews will be elaborated to illustrate how they can contribute as validity evidence.

Dimensionality and Science Assessment

As mentioned earlier, dimensionality analysis, DIF, and think-aloud interviews have been shown to be the most common approaches used to provide evidence in the formulation of the validity argument. Much of the research on the construct validity of large-scale science assessments has focused on test dimensionality. Test dimensionality is defined as the smallest number of "dimensions or statistical abilities required to fully describe all test-related differences among the examinees in the population" (Tate, 2002, p. 184). Knowledge of the latent dimensional structure can provide more meaningful information about test scores, and ultimately enhance the validity of the inferences made from the test scores (Ayala, et al., 2002; Childs & Oppler, 2000; Frenette & Bertrand, 2000; Hamilton et al., 1995; Nussbaum et al., 1997). The result of this could be that the inferences generated from test scores would systematically provide information about student's strengths and weaknesses in test performance. Dimensionality research can help answer questions such as "how many latent traits are being measured by a test overall?" and "is reporting student performance with a single score reasonable given the number of latent traits found to underlie the test?" These types of questions have led researchers to examine the dimensionality of tests of science achievement in an attempt to better understand the complex nature of students' cognitive skills in this area and how it interacts with measures of achievement, which may result in hypotheses about those attributes in students that causally influence the outcome of the measurement procedure (i.e., the test score) (Borsboom, 2005).

In order to establish whether unique dimensions of performance in science could be derived from the NELS: 88, Hamilton et al. (1995) examined the factor structure of this test for grade 8, 10 and 12 samples. The NELS: 88 assesses science, math, history, and reading of eighth graders. The NELS: 88 assessment is unique in that it is a longitudinal measure used again for the same grade 8 students when they are in grades 10 and 12. The Science test consists of 25 multiple choice items that assess a range of content and processes in the environmental, biological, and physical sciences. Some of the NELS science items measure factual knowledge, and problem solving and reasoning. From the 25 items that were administered to the grade 8 sample, 7 of the items were replaced to reflect curricular changes and difficulty at the 10th grade. At the 12th grade, 5 items that were common to the grade 8 and grade 10 and 1 solely to the grade 10 NELS: 88 administrations were replaced by new items (Nussbaum et al., 1997). The dimensional structure that emerged after subjecting the NELS: 88 science test samples for the 8th, 10th, and 12th grades to a full information factor analysis contained four (8th grade) and three correlated factors (10th and 12th grades). For the eighth grade NELS: 88 sample, factor analyses of the items yielded four factors (a) everyday science knowledge (ES), which included items that required knowledge that would be learned outside of the formal school setting and did not place great demands on school-acquired knowledge and understanding; (b) scientific reasoning (SR), which included items that required the manipulation of equations, interpretation of graphs, and hypothesis formation; (c) chemistry knowledge (CK), which included items that called for concepts such as mixtures, compounds, chemical change, and solubility but made few reasoning demands on students (this factor was determined primarily on subject matter); and (d) reasoning with knowledge (RK), which included items that required reasoning applied to formal science concepts, especially about science terms appearing in the multiple-choice response options (Hamilton et al., 1995). Hamilton et al. (1995) found that items clustered around content matter; however, obvious divisions between science domains were not recovered. A for a full description of the grade 8 factors that were recovered is provided in Appendix A. Leighton, Gokiert, and Cui (2007) suggested that although content is not entirely convincing when making claims about cognitive processing, it offers a preliminary point of examination because content is a key factor in reasoning and problem solving. The factor structure for the 10th grade sample included quantitative science (QS), spatial-mechanical reasoning (SM), and basic knowledge and reasoning (BKR). The 10th grade science items corresponded to distinct science achievement domains. Nussbaum et al. (1997) retrieved the same three factors, QS, SM, and BKR identified for the grade 10 sample for the 12th grade sample. They indicated that the QS factor found for both grades 10 and 12 samples is the combination of the CK and SR factors found for the grade 8 sample. They further found that factors that were slightly correlated at the 8th grade became more correlated and in fact indistinguishable at later grades, leading to a 3-factor model, most likely due to greater standardization of curriculum in high school. It should be noted, however, that the 4-factor model also met chi-square change criteria at the 10th grade; however, the factors were difficult to interpret, possessed weak loadings, and items had larger loadings on the other 3 factors (Nussbaum et al., 1997). Another possible explanation could be the small number of items that made up the NELS: 88 and subsequent item changes that were made at the grade 10 and grade 12 levels as contributing to the difference between the 4 factor and the 3 factor solutions. A later compilation of NELS: 88, TIMSS, and NAEP multiple choice items were tested in a confirmatory model and the three factors, QS, SM, and BKR, were retrieved (Ayala et al., 2002). This research suggests that the NELS: 88 was in fact multidimensional at all grade levels (Ayala et al., 2002; Hamilton et al., 1995; Nussbaum et al., 1997). In addition to Richard E. Snow pioneering work a number of researchers have found large-scale science assessments to be multidimensional (e.g., Ayala et al., 2002; Frenette & Bertrand, 2000; Hamilton et al., 1995; Leighton et al., 2007; Nussbaum et al., 1997).

Leighton et al. (2007) tested the 3- and 4-factor models, previously found by Hamilton et al. (1995) and Nussbaum et al. (1997), using the 1999 13-year-old and 16-year-old SAIP science samples. They also tested the following four additional models: item format (multiple choice and constructed response), test specifications (which included six content domains), and skill specifications (use, procedural, and conceptual). Finally, items were coded according to abductive and deductive reasoning as put forth in Lawson's (2005) hypothetico-deductive model. LISREL was used to test the six models in a confirmatory factor analysis. Root mean square residual (RMR), the adjusted goodness of fit index (AGFI), and the chi-square statistic were compared for best model data fit. The results indicated that the 4 factor model found in the NELS: 88 grade 8 sample and the 3 factor model found for the grade 10 and 12 students fit both the 13- and 16- year-old SAIP data well. Furthermore, item format, test specifications, skill specifications, and Lawson's hypothetico-deductive model also indicated good model data fit. Based on these results, Leighton et al. (2007) argued that it is not surprising that all models fit given that the SAIP items are complex and measure multiple content domains, cognitive processes, and item formats.

Although several basic approaches exist to determine how many factors to retain in factor analysis, it is not clear which approach is the most effective (Mislevy, 1986; Preacher & MacCallum, 2003). Beyond purely statistical considerations, arguably the most important thing to consider is whether the factors are interpretable (Gorsuch, 1983). Even though a number of exploratory and confirmatory approaches to the study of dimensionality exist, the literature focused on large-scale science assessment is largely influenced by the exploratory approaches (e.g., Hamilton et al., 1995; Leighton, Gokiert, & Cui, 2005). Results from exploratory analyses can be used, as a data-driven method, both to investigate whether a science assessment is measuring a multidimensional construct, and to guide the development of hypotheses about scientific reasoning. When conducting confirmatory analyses, as was demonstrated by Leighton et al. (2007), test specifications can act as a springboard for examining the dimensional structure of science content and skills. However, it has been argued that test specifications do not capture subtle psychological processes and therefore, may not explain the underlying cognitive structure of the data well in a confirmatory paradigm (Leighton et al., 2005; Norris et al., 2004). However, if test specifications are used as a blueprint from which test items are developed to measure the scientific knowledge and skills of students, then it could be assumed that these knowledge and skill areas would be elucidated in a confirmatory analysis. However, it appears as though model data fit is highly dependent upon the nature of the model to be confirmed. The lack of fit of a model derived from the test specifications calls in to question the methods and theory that test developers adhere to when developing test specifications .

That there is a lack of fit between test specifications and the data in the form of student responses in some cases is not surprising given that test specifications do not necessarily represent the cognitive processes students use to respond to test items (Norris et al., 2004). As a result, there has been a call for the use of cognitive models to better guide large-scale achievement assessment development (Embretson, 1999; Haladyna & Downing, 2004; NRC, 2001; Leighton, Gierl, & Hunka, 2004; Snow & Lohman, 1989). The National Research Council (2001) suggested that more meaningful inferences could be made about student knowledge and skills if they were tied to explicit theories of cognition and learning. Theories of scientific reasoning exist; however, these theories are conceptual in nature, have not been used when developing tests, and are rarely used to describe student performance based on test scores. The majority of dimensionality studies examine test data after the test has been administered, and attempts to match test score interpretation to existing theories of scientific reasoning occur after the data have been collected (Leighton et al., 2004; Leighton et al., 2005; Leighton et al., 2007; NRC, 2001). It would seem logical for a theory of scientific reasoning to guide test development and then be used to interpret test scores; however, it is often the case that tests are designed without a theoretical model in mind (Lane, 2004; Leighton et al., 2005; Leighton et al., 2007). “Retrofitting” data to existing theories of scientific reasoning, although well intentioned, may result in hypotheses about test score interpretation at best, and inappropriate model data fit at worst.

Differential Item Functioning and Science Assessment

Ferrara et al. (2004) suggested that the interpretation of "tests scores is valid only to the degree to which a test's component items are construct valid" (p. 1). When attempting to describe the dimensional structure of a test it is also important to identify and understand implicitly the dimensional structure of individual items. Roussos and Stout (1996a) define "dimension of an item" as "any substantive characteristic of an item that can affect the probability of a correct response on the item" (p. 356). When an item is found to measure multiple dimensions, this can result in differential item functioning (DIF) for groups of students. DIF occurs when two groups of examinees with equal ability do not have the same probability of answering the item correctly. It has been suggested that items that display DIF measure a primary dimension (the dimension the item is intended to measure) along with at least one secondary dimension, which was not intended to be measured by the item (e.g., Gierl, 2005; Messick, 1989; Roussos & Stout, 1996a). The secondary dimension(s) that are measured by the item can be representative of the construct or irrelevant to the construct being measured. Construct-irrelevant variance has been described as a form of systematic error that can affect the probability of an examinee answering an item correctly (Messick, 1989). When a construct-irrelevant (or nuisance) dimension is present on a test, examinees with lower ability on the nuisance dimension will likely score lower on the test than other examinees who are of equal ability on the dimension of interest, but who have higher ability on the nuisance dimension. On the other hand, an additional dimension (auxiliary) can be measured within a test or test item that is relevant to the construct. When a student possesses more of this secondary dimension they will perform better on the item resulting in *impact*.

According to the language of Shealy and Stout's (1993a) Multidimensional Model for DIF (MMD), if the DIF is caused by an auxiliary dimension it is considered *benign* (intentionally assessed which may reflect impact). Conversely, if a nuisance dimension causes the DIF it is considered *adverse* (unintentionally assessed which may reflect bias) (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Roussos & Stout, 1996a).

There are statistical methods for computing the probability that individuals of equal ability from different groups will answer an item correctly (Hambleton, Swaminathan, & Rogers, 1991). However, these methods do not always yield the same results when applied to the same data set (Clauser & Mazor, 1998). The most commonly used procedures include the Mantel-Haenszel (MH; Holland & Thayer, 1988), Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a), and Logistic Regression (LR; Swaminathan & Rogers, 1990). SIBTEST has proven to be more accurate in detecting DIF than both the MH and LR methods (Jiang & Stout, 1998; Gierl, Rogers, & Klinger, 1999; Puhan, 2003). SIBTEST is advantageous for the following reasons: SIBTEST uses original item response data when conducting DIF analyses, rather than parameter estimates which rely on strong assumptions about the procedures underlying item responses (Hambleton et al., 1991). When conducting DIF a matching subtest is important in the identification of DIF but is subject to error. The matching subtest may contain a few biased items; therefore, contaminating the matching subtest. Simulation studies, however, have demonstrated that the SIBTEST can tolerate a small to moderate amount of contamination in the matching subtest (Shealy & Stout, 1993a, 1993b). SIBTEST employs a regression estimate for the true score rather than the observed score to match students with the same ability, which is useful in controlling for

Type I error. SIBTEST also uses an iterative procedure, which involves repeated analysis and removal of items that are flagged for DIF. Items are analyzed successively until no further items are suspected to contain DIF. This subset of items, labelled the matching test (sometimes termed anchor or valid test), is considered to be DIF-free and is used as a basis of comparison for the suspect items on the test. SIBTEST includes a test of significance, can be run in bundle format, possesses reasonable statistical power, and performs well with relatively small samples (Clauser & Mazor, 1998; Gierl et al., 1999).

Once items have been identified as statistically significant for differential item functioning (DIF), the next step is to determine whether this difference is due to bias or impact. Bias, in the context of assessment, occurs when items on a test systematically advantage or disadvantage one group over another even when the groups have the same ability. This bias may result in the inconsistent selection and classification of students, which can have potential consequences if the nature of the selection and classification is high stakes (Moss, 1998). Conversely, an item displays impact if the identified DIF is due to construct-relevant aspects such as genuine knowledge, experience differences, or both. The capacity to determine whether the DIF is due to bias or impact is somewhat under-developed. Although, DIF analyses are a routine part of some large-scale assessment testing programs, less common are studies to understand the potential sources of DIF (Gierl et al., 2001). Some sources of DIF in large-scale assessments that have been explored include differences in item format (multiple choice vs. open-ended/constructed response), gender, translation, culture, and background experience (e.g., Ercikan et al., 2004; Gierl et al., 1999; Henderson, 1999; Lin, 2006). However, in the domain of science the most common sources of DIF explored are gender and item format.

Sources of DIF are often examined and identified through expert review together with statistical differential item functioning (DIF). Statistical analyses of items and review of items by experts have been used to justify the removal of potentially biased items from a test. Expert review of items is a substantive approach, which can be considered subjective, whereas statistical procedures can be considered as purely data driven. When combining the two approaches, substantive and statistical, it is often the case that the items flagged by each method do not always match (Gierl, 2005). Therefore, neither of these procedures used alone or together necessarily warrants removal of items from a test. One area where the limitation of substantively understanding DIF has been noted is gender differences. Specifically, gender differences in science assessment and achievement continue to be an area of interest to many researchers.

Gender differences in science assessment. Gender differences in large-scale assessment have been considered by many to be the most carefully examined aspect of test fairness (Ryan & DeMark, 2002). Maccoby and Jacklin's (1974) pioneering work on gender differences shaped the current research trend toward examining the accuracy of claims that males and females differ on verbal ability, quantitative ability, and spatial ability. Linn and Hyde (1989) conducted a meta-analysis and the findings suggested that gender differences in cognitive and psychosocial tasks are small and declining, are specific to cultural and situational contexts, and often reflect differences in course enrolment and training. Although the cognitive differences between males and females in science are reportedly small, women are still noticeably underrepresented in the science and mathematics occupational fields. The access and interest of women pursuing careers in the scientific field has slightly increased; however, wages for science related occupations still remain disparate among males and females (Johnson, 1996; Linn & Hyde, 1989; Penner, 2003; van Langen, Bosker, & Dekkers, 2006). Over the last few years, reform efforts in Canada and the United States have placed emphasis on improving science achievement for all students and, in particular, ensuring that females are given the necessary support and encouragement to pursue scientific studies and careers (Enman & Lupart, 2000; Hamilton, 1998).

Hedges and Nowell (1995) synthesized the results from several gender difference studies that explored gender differences in large-scale assessments among nationally representative samples. The impetus for their study was to investigate the hypothesis that substantial gender differences occur as a result of opportunity to learn. Overall, their analyses suggested that gender differences were small for most areas of achievement, with the exception of writing achievement, science achievement, and stereotypically male related occupations (Hedges & Nowell, 1995)[2]. Although some findings suggested that average gender differences in science were decreasing (Linn & Hyde, 1989), Hedges and Nowell (1995) found that across the 32-year period they examined, gender differences were relatively stable. Furthermore, Hedges and Nowell's results were not consistent with the hypothesis set forth by Linn and Hyde (1989) who suggested that gender differences are a result of opportunity to learn. Contrary to their hypothesis, Hedges and Nowell (1995) found substantial differences in writing performance, a skill that is taught to all students. Research in the area of gender differences in science is abundant; however, the findings are somewhat inconsistent.

---

[2] The stereotypically male related occupations were left largely unspecified throughout this study.

Research on gender differences, specific to the assessment of science, reveal trends in content and skill areas in which males and females differ (Beller & Gafni, 1996; Halpern 1997; Hamilton, 1998; Hedges & Nowell, 1995; Linn & Peterson, 1985). These trends are especially apparent in spatial ability, physical science, and earth and space science items, which reveal large male advantages. A statistically significant male advantage was found in Hedges and Nowell's (1995) study of the NELS: 88 multiple-choice science test. When Hedges and Nowell considered the dimensional structure found previously for the NELS: 88 (Ayala et al., 2002; Hamilton et al., 1995; Nussbaum et al., 1997), they attributed the difference to the performance on the spatial mechanical reasoning (SM) dimension. The spatial mechanical and quantitative reasoning (QR) dimensions consist primarily of items that could be classified as physical science items. The gender differences found on the QR dimension and the basic knowledge and reasoning (BKR) dimension were insignificant. Beller and Gafni (1996) analyzed the 1991 International Assessment of Educational Progress (IEAP) and found a significant male advantage on physical science and earth and space science items. A similar pattern of male advantage was found for fourth grade students on the Third International Math and Science Study (TIMSS); while males outperformed females on physical and earth science items, little difference was found between males and females for life and nature of science, or for environmental issues (Hamilton, 1998). Furthermore, the grade 8 results for the 1995 administration of the TIMSS demonstrated statistically significant mean gender differences for Canadian students in science with boys achieving higher performance than girls (Beaton et al., 1996).

Although gender differences are frequently found on spatial ability measures at both the item and test levels, it is unclear how spatial ability affects performance in science achievement. If spatial ability is a significant construct that should be measured by large-scale science tests of achievement and if boys tend to possess more of this skill are we ultimately disadvantaging female students by including items of this type? Or, is it that we are enabling males and females to demonstrate their strengths and weaknesses on that construct. Additional research suggests that males are more attracted to extracurricular activities and courses that establish and enhance spatial abilities (Hamilton, 1998); however, this hypothesis has not been fully investigated or empirically tested. Halpern (2004) suggested that females typically achieve higher grades in school, and possess the tendency to score much higher on tests that involve writing and that capture content that is similar to what they have learned in school. On the other hand, Halpern's results suggest that males receive higher scores on standardized tests of math and science that are not directly linked to school content, and that males show large advantages on visuospatial tasks that involve velocity judgments and navigation through three-dimensional space. More recently, Pope, Wentzel, Braden, and Anderson (2006) examined the relationship between gender and large-scale Alberta provincial achievement tests in grades 3, 6, and 9. They found that female students outperformed males in language-based tests (i.e., writing and reading); on the other hand, males outperformed females in mathematics and science.

Item format differences. Another source of DIF is item format. The curriculum reform efforts of the 1990s prompted test developers to consider the broader implementation of constructed-response assessments (Ryan & Demark, 2002). This push toward the use of CR formats in large-scale assessment raised several important validity concerns. Ryan and Demark (2002) highlighted some of these concerns by answering the following questions: “Do constructed response formats operationally define curriculum domains that include construct-related or construct-irrelevant sources of variation? Do constructed-response formats reflect verbal comprehension and production abilities that influence all assessments using this format? Are students with superior verbal skills advantaged by the use of constructed-response formats regardless of the content area being measured?” (p. 69). The likelihood that male and female performance diverges on item format (e.g., MC and CR) has resulted in several studies into this possible source of test bias (e.g., Klein et al., 1997; Resnick & Resnick, 1992). The general trends have indicated that males tend to perform better than females on MC items in science whereas females perform better on CR tasks in science (Resnick & Resnick, 1992). The possibility that MC and CR tasks measure different cognitive skills may explain, in part, why males and females perform differently across these tasks. If MC and CR are measuring different aspects of achievement, an interaction between item format and gender might be expected.

Stumpf and Stanley (1996) suggested that females experience performance advantages when scores depend on language usage, therefore, resulting in a female advantage on CR tasks. Although Stumpf and Stanley were tentative about this suggestion, findings from other studies on gender differences have supported their explanation of item format differences (Henderson, 1999; Klein et al., 1997). Klein et al. (1997) furthered the “language usage” hypothesis by demonstrating that females generally performed better than males on hands-on science tasks that required attention to detail and reading. On the other hand, males outperformed females on items that required inferences and prediction. Beller and Gafni (2000) suggested that women's verbal abilities may be better illuminated in constructed response items than in multiple-choice items. They further suggested that writing ability may also play a role in the differential performance of females on constructed response items. In addition, they suggested that males may perform better on MC items as they take more risks in responding. A comprehensive review of gender and fair assessment conducted by Willingham and Cole (1997) led to the conclusion that although females tended to perform better on CR formats than on MC formats, this effect was not consistent, as many studies also demonstrated that females can perform as well on MC items. They also found that item format differences between males and females in mathematics, language, and literature did not occur as frequently as item format differences between males and females in science. In conclusion, the domain of science achievement presents an especially unique challenge, which may necessitate diverse sources of validity evidence.

Verbal Reports and Science

Small-scale interview studies offer one method that can be used to shed light on the gender differences associated with test item performance (Hamilton, 1999; Ercikan et al., 2004). Beyond aiding in hypothesis generation about the source of group differences in science, interview data can also help yield hypotheses about the underlying knowledge and skills measured by tests. Combining substantive evidence with psychometric methods could yield convergent evidence to support the inferences that are made from test results (Hamilton et al., 1997). The interest in student performance, which goes beyond simple right and wrong response patterns, has increased the need for data that outline cognitive processes (Russo, Johnson, & Stephens, 1989). Think-aloud verbal protocols in which students are asked to verbally report their thoughts as they work through specified tasks have proven useful in examining the underlying cognitive skills that students employ in problem solving (Ercikan et al., 2004; Ericsson & Simon, 1993; Hamilton et al., 1997; Norris, 1990; Norris et al., 2004). Think-aloud methods offer one way to uncover the substantive nature of dimensions at both the test level and item level (Hamilton et al., 1997; Leighton, 2004; NRC, 2001). The National Research Council (2001) suggested that the validity of inferences drawn from test performance can be improved when information is gathered about the specific knowledge and skills students actually use during test performance. The common approach in determining the knowledge and skills measured by tests is to consult with content experts, test developers, and psychometricians. An inherent limitation to this approach is that content experts typically possess very different problem solving skills than students. Therefore, the hypotheses they generate or inferences they make about student performance may be misinformed (Leighton, 2004; Norris et al., 2004). Think-aloud verbal reports offer an alternative way to support statistical investigations by allowing researchers to examine more precisely the scientific reasoning skills that students employ (Ercikan et al., 2004; Ericsson & Simon, 1993; Hamilton et al., 1997; Hamilton, 1998; Leighton, 2004; Norris et al., 2004) as they solve science tasks.

Ericsson and Simon (1993) reviewed over 50 studies and concluded that when verbal reports are collected under specific conditions they can provide important and reliable information about the cognitive processes that students engage while solving tasks. In their review, these researchers made a distinction between concurrent verbalizations and retrospective verbalizations. Concurrent verbalizations involve students verbalizing the information that they are attending to while solving the task. Retrospective verbalizations occur after the task has been completed and the student is asked to recall their thought processes. The utilization of verbal reports is regularly found in the psychological literature; less common are studies that have employed this procedure when examining educational tasks (Ayala et al., 2002; Ferrara et al., 2004; Gierl, 1997; Hamilton et al., 1997; Leighton & Gokiert, 2005; Lin, 2006; Rogers & Bateson, 1990; Yepes-Baraya, 1996).

Identifying knowledge and skills. The collection of verbal reports is useful because they can yield evidence for construct validity by elucidating student cognitive processing on achievement tasks (Leighton, 2004; Norris et al., 2004). Studies that have utilized this method have yielded information about the constructs that are measured by test items and potential explanations for student performance differences (e.g., Ercikan et al., 2004; Hamilton et al., 1997; Lin, 2006). Baxter and Glaser (1998) suggested an analytic framework for evaluating test constructs in science assessments by examining the relationship between verbal protocols, observation of student performance, and scoring criteria. They used the verbal protocols of a small number of subjects to empirically assess the tasks. Baxter and Glaser proposed that students that possess competency in problem solving and engage in complex tasks have integrated knowledge (this type of knowledge allows students to generate inferences with what they know) and usable knowledge (this is knowing what knowledge to apply in different situations) vs. fragmented knowledge (this is not knowing when to apply knowledge of conceptual or procedural skills to given situations). Hamilton et al. (1997) used a small-scale interview study to aid in the interpretation of factors that were uncovered in the dimensionality analyses of the NELS: 88 science assessment. Forty-one high school students were asked to complete 16 multiple-choice items selected from the 10th grade NELS: 88 science test. These 16 items represented the QS, SM, and BKR dimensions found by Hamilton et al. (1995). Four additional constructed response questions were included, three science items and one mathematics item with science content. From this study, the researchers concluded that small-scale interviews could be used to enhance and support dimension interpretation in order to define the construct more clearly. The interviews proved helpful in interpreting items that possessed inconsistent factor loadings and provided valuable insights into students' cognitive processes and beliefs about their problem solving that were not evident to researchers from simply reading the items. They argued that multiple test formats and methods of analysis are critical when exploring the complexity of students' cognitive processing and interpreting test results (Hamilton et al., 1997).

Think-aloud interviews have also proven useful for identifying how students comprehend test items, the knowledge and skills they use in problem solving, and how ambiguities in test items can thwart student problem solving. Leighton and Gokiert (2005) investigated how students comprehend large-scale science test items and the knowledge and skills they report as useful when solving these types of items. To begin, they examined 30 test items to identify ambiguities and found that (a) words and phrases, (b) background context, and (c) structural features had the potential to derail student problem solving. Fifty-four students (30 grade 8 students and 24 grade 11 students) were requested to think-aloud concurrently as they solved item sets consisting of five items. Following the concurrent portion of the interview, students were asked retrospective questions aimed at probing their metacognitive knowledge (i.e., beliefs about their problem solving, including their evaluation of the task). The findings indicated that although students had a difficult time verbalizing their thoughts concurrently, they were able to describe retrospectively their comprehension of the items and potential ambiguities in the items. The researchers argued that ambiguous words, phrases, or features in an item could result in information irrelevant to the construct that could, in turn, impact a student's ability to demonstrate their knowledge and skills properly (Leighton & Gokiert, 2005). Ferrara et al. (2004) examined the alignment between "content area knowledge, content area skills, broader cognitive processes, and response strategies (Knowledge, Skills, Processes, and Strategies: KSPS)" that test developers purport to measure (intended) and the actual KSPS that students apply when answering test items (observed) on a grade 6 state assessment. They suggested that alignment analysis can be used to investigate the construct validity of items. In order to address the alignment between intended and observed KSPS, the researchers compared empirically derived coding categories of intended KSPS and compared these categories with think-aloud data, which identified the observed KSPS outcomes. As evidenced by the think-aloud protocols, the results indicated that simple words and phrases in the items could significantly impact the ways in which examinees responded to the tasks, thought about the tasks, and ultimately selected their final responses. Their results further suggested that these subtleties are overlooked by item writers and reviewers as they are not specifically trained in the association between item targets and student cognitive processing. Explicit knowledge of the alignment between item targets and cognitive processing would improve the construct validity of items and ultimately the tests. If the goal is to make valid inferences about student performance, it is imperative to examine the underlying knowledge and skills students bring to bear on tests of achievement.

Identifying group differences. Interview studies that have been conducted on science assessments reveal that extracurricular activities may play a significant role in boys' superior performance on spatial mechanical reasoning items and physical science test items (e.g., Hamilton, 1999). Involvement in community extracurricular activities specific to science can enhance performance in this area for both males and females; however males are more likely to engage in activities outside of school that would develop math and science skills (Hamilton, 1999; Linn & Hyde, 1989). Some of these activities include dismantling mechanical or electrical objects, assisting with car maintenance, playing with constructional or electrical toys, and reading about science (Johnson, 1987). These findings suggest that the more educators, parents and community members encourage females to engage in extracurricular activities and course taking patterns related to science, the gap between male and female performance on those tasks which require spatial mechanical abilities may be bridged.

Through the use of statistical DIF analyses and small-scale interviews, Hamilton (1999) found gender differences on NELS:88 science items that required visualization and items that required knowledge and skills obtained outside of the educational setting (i.e., car maintenance and dismantling mechanical objects). If males possess stronger visualization skills and are more apt to use them in solving science items, this could explain why boys outperform girls in spatial reasoning. To fully appreciate how the multifaceted associations between format, content, and cognitive processes affect the performance of different groups of students, the investigation of possible contributing item features needs to be approached cautiously (Hamilton, 1999).

Ercikan et al. (2004) combined DIF analysis, expert review, and think-aloud protocols to investigate the differences between English and French students on the SAIP mathematics and science tests. After items were identified as displaying statistical DIF, assessment experts bilingual in French and English examined the items to generate hypotheses about the sources of DIF. Grade 7 and 8 students (36 English speaking and 12 French speaking) were asked a set of questions after they solved each of the 20 math and science items that were identified as displaying DIF between the two language groups. The questions were aimed at gauging the students' understanding of the intent of the item, the steps they took to solve the item, the reasons for selecting the answers that they did, and what parts of the item helped or hindered their problem solving. The test administrators were instructed to ask those questions that had not been spontaneously reported by the participants' think-aloud process. The results indicated that the hypothesis set forth by the assessment experts were consistent with student reports on six of the items. For the remaining 14 items, the protocols did not match the hypothesis put forth by the assessment experts. Ercikan et al. (2004) concluded that think-aloud protocol analysis is a promising technique when determining sources of DIF and should be used as a complimentary method to statistical DIF analysis like judgmental reviews. They also suggested that the student's think-aloud reports resulted in additional hypotheses about why the DIF was occurring that were not identified by the expert judges.

Similar to Ercikan et al. (2004), Lin (2006) conducted a study examining French and English translation/adaptation issues within the domains of math and social studies.

Grade 9 achievement tests were developed in math and social studies using the simultaneous test development approach. First, six accredited translators reviewed the items for comparability in meaning and form. If items were found to be different in meaning or form, the translators had to justify their reasons for making this claim. Second, 26 mathematics items and 40 social studies items were examined for DIF. Finally, concurrent and retrospective think-aloud data were collected from a sample of 24 English speaking and 39 French Immersion students while they solved a selection of DIF and non-DIF items from the mathematics and social studies tests. The student responses were used to highlight whether students found cues in the items that were helpful or made things more difficult. Results indicated that the interview data revealed no evidence for adaptation as a source of DIF. However, it was noted that French immersion students' lack of proficiency in French could be contributing to the differential item functioning found.

Summary of Literature Reviewed

In light of the increased use of large-scale science assessments and given the potential outcomes of test scores derived from such tests, understanding the validity of these tests has become more important. One means of exploring validity of science tests has been the assessment of dimensionality. Research has demonstrated that dimensionality analysis can enhance the validity of the inferences made from test scores by providing a better understanding of the latent dimensional structure (the knowledge and skills) of the test (Ayala et al., 2002; Frenette & Bertrand, 2000; Hamilton et al., 1995; Leighton et al., 2007; Nussbaum et al., 1997). For example, it has been demonstrated that the NELS: 88 science test possessed unique dimensions for the grade

8, 10, and 12 samples. Furthermore, a dimensionality investigation of the SAIP Science Assessment for grades 8 and 11 confirmed that the test could be explained by more than one dimension at both age groups (Leighton et al., 2007).

Not only is it important to understand the dimensional structure of the test, it is also important to identify and understand the dimensional structure of individual items. When items are found to measure multiple dimensions, this can result in differential item functioning (DIF). One of the most carefully examined forms of DIF in large-scale science assessments is the performance differences between males and females (Ryan & DeMark, 2002). Although some findings suggested that average gender differences in science were decreasing (Linn & Hyde, 1989), Hedges and Nowell (1995) found that across the 32-year period they examined, gender differences were relatively stable. The research on gender differences reveals that males tend to perform better on those tasks that involve spatial abilities and visualization (Halpern 1997; Hamilton, 1998; Hedges & Nowell, 1995). It has been hypothesized that extracurricular activities related to science and mathematics may enhance abilities on science tasks that require spatial visualization. It has been found that female students outperform males in language-based tests (i.e., writing and reading) (Halpern, 2004; Pope et al., 2006). To date, however, gender research for the SAIP Science Assessment is somewhat limited. Although dimensionality and DIF analyses are a routine part of large-scale testing programs, very few studies have gone beyond the statistical results to substantively document why the DIF may be occurring (Gierl et al., 2001; Gierl, 2005) and what impact it may have on male and female science achievement overall. Throughout the literature it has been demonstrated that small scale interviews of students verbally reporting their thought processes as they solve test items can shed light on the gender differences associated with test item performance (Hamilton, 1999; Ercikan et al., 2004). Although new to the field of educational measurement, verbal reports have shown promise in the psychological literature for examining the underlying cognitive skills that students employ in problem solving (Ericsson & Simon, 1993; Chi, 1997), lending credibility to both DIF and dimensionality results (Hamilton, 1999; Hamilton et al., 1997; Norris et al., 2004).

The intent of this study was to coordinate three forms of validity evidence, dimensionality analyses, DIF, and think aloud interview data, to support the inferences that are drawn from the SAIP Science Assessment. This type of evidence has the potential to arm developers of large-scale science assessments with more knowledge about the potential biases inherent in science test items, and the cognitive differences that male and female students possess. With this knowledge, test developers would be in a better position to make evidence-based decisions about which items to retain, modify, or discard and how to report overall student performance.

Chapter Three: Methods and Results: Psychometric Approaches

To investigate whether the grade 8 English sample data on the SAIP Science Assessment fit the 4-factor model put forth by Hamilton et al. (1995), confirmatory factor analysis using LISREL 8.53 was conducted (Jöreskog, & Sörbom, 2002). The SAIP science items were also investigated for differential item functioning (DIF) using the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a). The focal group consisted of the female sample and the reference group comprised the male sample that wrote the SAIP Science Assessment. Finally, verbal reports based on a sample of SAIP questions identified as displaying gender DIF were collected for male and female grade 8 students using think-aloud methods (Ericsson & Simon, 1993; Leighton & Gokiert, 2005). For ease of presentation and interpretation, the methods and results will be presented sequentially for each separate test (AB and AC tests – to be explained in the next section) for each of the methodologies used in the study. First, the SAIP science assessment will be described in detail, followed by a description of the grade 8 sample of students that completed the AB and AC tests. Second, the confirmatory factor analysis method and results will be presented for each of the separate tests (AB and AC), followed by the methods and results for the DIF on each of the tests. The think-aloud interview methodology followed by the results of aggregated student reports on each item are provided in the next chapter.

Data

SAIP Science Assessment

The current study involved the written portion of the Science assessment data collected from 13-year-old students across Canada in the 2004 administration of the

School Achievement Indicators Program (SAIP; CMEC, 2004). Although the SAIP is administered to both grade 8 (13-year-old) and grade 11 (16-year-old) students, the present study was restricted to the grade 8 sample. While the SAIP data were available for the 16-year-old students, the participation rate for the interview portion of the study was insufficient (2 participants) to complete each of the three approaches to construct validity (dimensionality, DIF, and think-aloud interviews).

The SAIP Science Assessment utilized a two-stage testing procedure by which students write an initial 12-item routing test (Test A) designed to assign them to a second stage test. The routing test consisted of both multiple-choice (10 items) and constructed response (2 items) items that were targeted to level three, which is of moderate difficulty. Students who received a score less than 8 out of 12 were routed to the second-stage B test, while students who received a score of 8 or greater were routed to the second-stage C test. The items in the SAIP Science Assessment were equally distributed across five ability levels (1 = low ability while 5 = high ability) across the three forms (A, B, and C). Test B contained 71 items, with the majority of items at levels 1 (25 items), 2 (26 items), and 3 (14 items) and a smaller number at levels 4 (4 items) and 5 (2 items). Test C contained 65 items at levels 3 (13 items), 4 (26 items), and 5 (26 items). Like the routing test, both the B and C tests contained multiple-choice (39 items and 38 items) items and constructed response (32 items and 27 items) items. Nineteen items were common to both the B and C tests, these items are primarily targeted to the third ability level with a few items from both the fourth and fifth ability levels. The test items included in the routing test and subsequent B and C tests represented six broad science content domains: (a) *knowledge and concepts of science*, included *biology* (life sciences), *chemistry* (physical sciences), *earth and space*, and *physics* (physical sciences) (b) *nature of science*, and (c) *relationship of science to technology and societal issues*. The knowledge and concepts of science more specifically measured within *biology*, *chemistry*, *earth* and *physics* include: (1) matter has structure and there are interactions among its components; (2) life forms interact within environments in ways that reflect their uniqueness, diversity, genetic continuity, and changing nature; (3) basic gravitational and electromagnetic forces result in the conservation of mass, energy, momentum, and charge; and (4) earth and the physical universe exhibit form, structure, and processes of change. Within the *nature of science*, the nature of scientific knowledge and the process by which that knowledge develops was measured. Finally, *the relationship of science to technology and societal issues* measured the specific skills associated with how performance in science involved an understanding of associations between science, technology, and society. Based on a student's performance on the B or C test items, they were assigned to one of six performance levels ranging from 0 = low ability to 5 = high ability. Students with a performance score below 3 demonstrated lower level science achievement in the measured domains, while students scoring 3 or higher demonstrated relative strengths in the measured science domains. The distribution of items across content and skill areas is provided in Appendix B.

Sample

The original English data set of grade 8 students who completed the SAIP Science assessment contained 10,096 response vectors. For analysis purposes, the data were divided to produce the AB and AC test data sets. Inspection of the two data sets revealed that some students had only completed one portion of the test. For example, a student may have completed the 12-item routing test, but did not complete either of the B or C tests. The responses for these students were deleted from the data files. The final sample sizes were 4,307 for test B and 4,199 for test C. The number of students, broken down by test and gender, are reported in Table 1. As shown, the tests are designated AB and AC to convey that each student completed the routing test A and the second stage test to which they were assigned.

Table 1

Sample of Grade 8 Students that Wrote the 2004 SAIP Science Assessment

	AB Test	AC Test	Total
Combined	4,307	4,199	8,506
Male	2,018	2,232	4,250
Female	2,289	1,967	4,256

Confirmatory Factor Analysis

The 4-factor Model (ES, SR, CK, and RK)

The first research question - Does the 4-factor model previously found by Hamilton et al. (1995) for the NELS: 88 grade 8 sample fit the SAIP 2004 Grade 8 sample data? If not, what is the dimensional structure of the 2004 version of the SAIP Science Assessment? - was answered by conducting a confirmatory linear factor analysis using LISREL 8.53 (Jorskog & Sorbom, 2002)[3]. Prior to conducting the analysis, the items on both the AB and AC tests were sorted according to the 4-factor model that included everyday or elementary science (ES), scientific reasoning (SR), chemistry knowledge (CK), and reasoning with knowledge (RK). In Leighton et al.'s (2007) study the 1999 SAIP items had been reliably categorized, using the factor descriptions put forth in Hamilton et al. (1995) and Nussbaum et al. (1997), by two pre-service teachers with specialization in science. These item categorizations were used in the present study (see Leighton et al., 2007 for an in-depth discussion of the item coding and inter-rater agreement). In consultation with a representative from CMEC, it was determined that from the 1999 SAIP Science Assessment to the 2004 version, 81 items were unchanged, 36 items received slight wording modifications to improve clarity, and 12 new items had been added (P. Brochu, personal communication, June 11, 2007). After all 129 items on the 2004 SAIP had been coded according to the 4-factor model, the 36 changed and 12 new items were re-evaluated for accuracy in coding by the researcher. The results of the coding for the AB and AC tests are reported in Appendix C. As shown, there were 30 ES items, 35 SR items, 9 CK items, and nine RK items for the AB test. For the AC test, there were nine ES items, 45 SR items, 20 CK items, and three RK items.

---

[3] The confirmatory factor analysis was conducted for both the male and female samples on each test (AB and AC) and the results indicated that the 4-factor model fit.

AB and AC results. The results of the confirmatory factor analysis for both tests are presented in Table 2. A number of goodness-of-fit indices exist to assess the fit of confirmatory factor analytic models; however, there is much debate in the literature surrounding which is the best index of model fit (Bollen & Long, 1993; McDonald & Marsh, 1990). For this reason, four indices were used to assess the model fit for the 4-factor model. The first was the Root Mean Square Error of Approximation (RMSEA) that adjusts for parsimony by assessing the discrepancy per degree of freedom in the model. The second index considered was the Root Mean Squared Residual (RMR), which is the root mean of squared residual when comparing the observed covariances fitted and the hypothesized covariances. RMSEA and RMR values of 0.05 or less indicate close fit of a model, while values of 0.08 reflect reasonable fit of a model (Browne & Cudeck, 1993; Reise, Widaman, & Pugh, 1993). The third index was the Adjusted Goodness of Fit Index (AGFI), which adjusts for the degrees of freedom of the model by substituting the total sum of squares with the mean squares; an AGFI value of at least 0.90 indicates good fit. The final index considered was the chi-square statistic, which indicates whether the restrictive hypothesis tested can be rejected. When using the chi-square statistic, a model has acceptable fit if the discrepancy between the variance-covariance matrix generated by the original data and by the hypothesized model is small, ultimately resulting in a nonsignificant chi-square. It should be noted that the chi-square statistic is sensitive to large sample sizes and can result in a statistically significant difference (Gierl & Rogers, 1996). Inspection of the values of the four fit indices reported in Table 2 indicated that the 4-factor model fit the data well for both the AB and AC tests. The RMSEA and RMR indices are below 0.05, and the AGFI is above 0.90. The chi-square statistic is significant at the 0.05 level of significance; however, given the large degrees of freedom this is not surprising (Gierl & Rogers, 1996).

Table 2

CFA of Hamilton et al.'s (1995) 4-factor Model Applied to SAIP Tests AB and AC

		LISREL Index				
Data	Number of Items	N	RMSEA	RMR	AGFI	X^2 *(df)*
AB	83	4307	0.02886	0.005773	0.9175	10812.01 (3314)
AC	77	4199	0.01528	0.003875	0.9644	5664.61 (2843)

Differential Item Functioning (DIF)

DIF Method

The Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a) was used to answer the second research question: do the performances of male and female students on the SAIP items systematically differ? SIBTEST is a statistical procedure that can be used to detect and estimate DIF on a given test. According to the Shealy and Stout model (1993a), an item is likely to display DIF if the item measures a secondary dimension along with the primary dimension the item was intended to measure. Camilli and Shepard (1994) and Roussos and Stout (1996a) suggested that when an item displays DIF it is because the item is measuring the primary dimension assessed by the full set of items and a secondary dimension, creating multidimensionality within a test. The secondary dimension may measure construct relevant information and be termed an auxiliary dimension, or measure construct-irrelevant variance and be termed a nuisance dimension. The purpose of SIBTEST, therefore, is to determine the difference between the probabilities of two groups, with equal ability on a latent trait, selecting a correct response. The reference group is typically considered the advantaged group while the focal group is typically considered the disadvantaged group. Although it is not clear which group would be considered advantaged or disadvantaged, for the purposes of this study the reference group included males and the focal group included females. The amount of DIF present in an item is denoted as $\hat{\beta}_{UNI}$, a parameter estimate with a standard normal distribution with mean of 0 and standard deviation of 1. If a statistically significant value of $\hat{\beta}_{UNI}$ is found to be positive, this indicates DIF against the focal group (i.e., the disadvantaged group is less likely to answer the item correctly); conversely, a negative value of $\hat{\beta}_{UNI}$ indicates DIF against the reference group (i.e., the advantaged group is less likely to answer the item correctly). Guidelines for determining the degree of DIF present have been provided by the Educational Testing Service (ETS; Zwick & Ercikan, 1989) and adopted by Roussos and Stout (1996b, p. 218, p. 220) and are outlined below:

No DIF: Null hypothesis is not rejected and $|\ \hat{\beta}_{UNI}\ | = 0$,

Negligible or A-level DIF: Null hypothesis is rejected and $|\ \hat{\beta}_{UNI}\ | < 0.059$,

Moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\ \hat{\beta}_{UNI}\ | < 0.088$,

Large or C-level DIF: Null hypothesis is rejected and $|\ \hat{\beta}_{UNI}\ | \geq 0.088$.

A comprehensive description of the technical aspects of SIBTEST can be found in Shealy and Stout (1993a).

DIF Results

For the present study, both the AB and AC tests were subjected to an item DIF analysis using SIBTEST. The overall SIBTEST results for the AB and AC tests are summarized in Table 3. Twenty-four out of the 83 (28.9%) AB test items displayed DIF, while 16 of the 77 (20.7%) AC test items displayed DIF.

AB results. Thirteen of the 24 DIF AB items possessed large or C level DIF ($\hat{\beta}_{UNI} \geq 0.088$) and, of the 13 items, 6 favoured males and 7 favoured females. The remaining 11 items displayed moderate or B level DIF ($0.059 \leq \hat{\beta}_{UNI} < 0.088$), 8 of which favoured males and three of which favoured females.

AC results. Seven of the 16 AC DIF items possessed large DIF ($\hat{\beta}_{UNI} \geq 0.088$) and of these, 3 items favoured males and 4 items favoured females. The remaining 9 DIF items displayed moderate DIF ($0.059 \leq \hat{\beta}_{UNI} < 0.088$), with 4 favouring males and 5 favouring females.

Table 3

SIBTEST results for the AB and AC Tests

	AB			AC		
	B-level	C-level	Total	B-level	C-level	Total
No. of DIF Items	11	13	24	9	7	16
No. of items favouring males	8	6	14	4	3	7
No. of items favouring females	3	7	10	5	4	9
Percentage of DIF items (%)	13	16	29	12	9	21

DIF and factor structure. As shown in Table 4, of the 24 DIF AB items, 8 items loaded on the everyday or elementary science (ES) factor, 14 items loaded on the scientific reasoning (SR) factor, no items loaded on the chemistry knowledge (CK) factor, and 2 items loaded on the reasoning with knowledge (RK) factor. Of the 8 DIF items that loaded on the ES factor, 5 favoured males (2 multiple-choice items and 3 constructed response items) and 3 favoured females (1 MC item and 2 CR items). Of the 14 items that loaded on the SR factor, 8 favoured males (6 MC items and 2 CR items) and 6 favoured females (4 MC items and 2 CR items). Finally, 1 DIF item favoured males and 1 DIF item favoured females on the RK factor; both items were MC items.

Table 4

AB DIF items by gender, factor loading, and format

DIF Items	Gender	Factor	Format
3	F	RK	MC
10	F	SR	MC
19	F	ES	MC
20	F	SR	MC
30	F	ES	CR
31	F	SR	CR
34	F	SR	MC
35	F	SR	CR
42	F	ES	CR
82	F	SR	MC
9	M	SR	MC
12	M	SR	MC
16	M	ES	MC
27	M	SR	CR
37	M	ES	MC
38	M	SR	MC
46	M	ES	CR
56	M	RK	MC
57	M	SR	MC
62	M	SR	MC
65	M	SR	MC
67	M	SR	CR
75	M	ES	CR
79	M	ES	CR

As shown in Table 5, of the 16 AC items displaying DIF, 2 items fell on the ES factor, 11 items fell on the SR factor, 2 items loaded on the CK factor, and 1 item loaded on the RK factor. All of the items that fell on the ES and CK factors were found to favour males and both CK items were MC while both ES items were CR. On the other hand, the RK factor DIF item was MC and in favour of females. Of the remaining 11 SR DIF items, 8 favoured females (6 MC items and 2 CR items) and 3 favoured males all of which were MC items. A complete listing of the items together with their factor loading, degree of DIF, format, and gender favoured are presented in Appendix D.

Table 5

AC DIF items by gender, factor loading, and format

DIF Items	Gender	Factor	Format
3	F	RK	MC
66	F	SR	MC
72	F	SR	MC
78	F	SR	CR
82	F	SR	MC
91	F	SR	CR
99	F	SR	MC
106	F	SR	MC
122	F	SR	MC
5	M	CK	MC
9	M	SR	MC
12	M	SR	MC
75	M	ES	CR
79	M	ES	CR
90	M	SR	MC
95	M	CK	MC

Selection of DIF items for think-aloud. Taking into account the results of the confirmatory factor analyses for the AB and AC tests, items that possessed statistically significant DIF and that were at B and C levels were reviewed for possible use in the interview portion of the study. It was determined that 12 items was a reasonable number of items for students to complete during a 45-minute interview so as not to infringe on class time. The following criteria were used to select items for the interview study: item representation from the 4-factors (ES, SR, CK, and SR); balanced item format representation (multiple choice and constructed response); equal gender DIF representation (female and male DIF items); items at differing difficulty levels (level one through five); and at least some items that overlapped between both tests. The 12 selected items are displayed in terms of these criteria in Table 6. As shown, 6 items favoured females while 6 items favoured males. Of the 12 items selected, 4 items represented everyday or elementary science (ES), 5 items represented scientific reasoning (SR), 1 item represented chemistry knowledge (CK), and 2 items represented reasoning with knowledge (RK). Five items were included in both the AB and AC tests, 4 items in the AB test, and 3 items in the AC test. Although not entirely equal, the 12 items represented item format (8 MC items and 4 CR items), difficulty level, and SAIP content domains.

Table 6

Items selected for protocol analysis

Item	Gender Favoured	Test	Format	Difficulty	SAIP Content	Factors
3	Female	AB /AC	MC	3	Knowledge of Physics	RK
9	Male	AB /AC	MC	3	Knowledge of Earth	SR
12	Male	AB /AC	MC	3	Knowledge of Earth	SR
30	Female	AB	CR	1	Nature of Science	ES
42	Female	AB	CR	2	Knowledge of Chemistry	ES
46	Male	AB	CR	1	Science, technology, and society	ES
56	Male	AB	MC	2	Knowledge of Earth	RK
79	Male	AB /AC	CR	3	Knowledge of Physics	ES
82	Female	AB /AC	MC	3	Nature of Science	SR
95	Male	AC	MC	5	Knowledge of Physics	CK
106	Female	AC	MC	4	Nature of Science	SR
122	Female	AC	MC	4	Science, technology, & S	SR

Items 3, 9, 12, 79, and 82 were selected because they represented 5 DIF items that were found to overlap between both the AB and AC tests. Of these 5 DIF items, 3 loaded on the scientific reasoning factor, 1 belonged to the reasoning with knowledge factor, and 1 loaded on the everyday or elementary science factor. Of the remaining 4 items from the AB test, items 30 and 42 belonged to the first factor everyday or elementary science. Of the remaining 3 items from the AC test, item 95 loaded on the chemistry knowledge factor, and items 106 and 122 belonged to the scientific reasoning factor. Although the 12

DIF items selected for the think-aloud portion of the study do not equally represent the 4-factors, the number of DIF items reflected the ratio of number of items that loaded on each factor. The two largest numbers of DIF items were selected from the ES and SR factors, which possessed the greatest number of items on both the AB and AC tests. Furthermore, the RK and CK factors had fewer items coded; consequently only 3 DIF items were selected for this factor for the think-aloud study.

Chapter Four: Method and Results: Verbal Reports

Think-Aloud Verbal Reports

A structured interview was used to probe students' cognitive processing and meta-cognitive knowledge as they solved the 12 selected test items (Ericsson & Simon, 1993) in an attempt to answer the third research question: if performance differences are found, can interview data of male and female grade 8 students aid in the generation of hypotheses about the underlying knowledge and skills that are measured by these items? The purpose was to gain further understanding of the source of the DIF found and to relate the student verbal reports to the factor structure identified in the confirmatory factor analysis. This process was used to determine if the DIF was related to real ability differences on the construct(s) of science achievement or to some form of construct-irrelevant variance.

Participants

A sample of Grade 8 students was recruited from a public junior high school in a major metropolitan area in Western Canada. In line with the goal of verbal reporting – to explicate the thinking processes of students as they solve items – the school from which the participants were drawn had a reported focus on developing metacognition and higher order thinking. A total of 50 informed consent forms were sent home with the grade 8 students that showed an interest in participating in the study. For a copy of the informed consent see Appendix E. Of the 50 informed consent forms taken home, 16 parents returned their forms, all of whom gave permission for their child to participate and for the release of the child's grades in language arts and science to the researcher. Nine of the students were male and seven were female. The mean science grade was 79.4% (*SD*

5.99%), with a range from 67 to 93%. The mean language arts grade was 78.2% (*SD* 4.23%), with a range from 66 to 85%. The female sample possessed slightly higher grades than the male sample in both science and language arts. The mean for females in science was 81.9% (*SD* 6.20%) and in language arts was 80.7% (*SD* 3.25%). The corresponding means for males were 77.4% (*SD* 5.36%) and 76.2% (*SD* 4.05%).

Interview

The 45-minute interviews were conducted over a two week span by the researcher and a graduate level research assistant who was trained by the researcher in think-aloud methodology. The research assistant was required to sign a confidentiality agreement (see Appendix F). At the beginning of the interview students were reminded of the purpose of the study and that they were free to exercise their right to opt out of the study without penalty. Students were then provided with an opportunity to practice the think-aloud process with one question (how many windows are there in your living room?). The practice question was presented to the students verbally while all remaining 12 items that were in the test booklet were presented in paper format. As the students solved the practice question, they were asked to think-aloud concurrently. After providing an answer to the question they were asked to retrospectively describe their thought processes. After it was determined that students were capable and comfortable thinking aloud, they were asked to open the test booklet containing the 12 DIF items and begin with the first item. Students were asked to think out loud as they solved each item. If students stopped talking during the concurrent portion of the interview, the interviewers encouraged students after 5 to 10 seconds to "please keep talking." After the students solved an item they were asked the following three retrospective questions aimed at probing their meta-cognitive knowledge of problem solving:

1. *Now tell me all that you can remember about how you solved this question*
2. *Did you find any parts of this question confusing? If so,*
   a. *What parts did you find confusing?*
   b. *Why were they confusing?*
3. *Did you find any parts of the question helpful in answering the question? If so,*
   a. *What parts did you find helpful?*
   b. *How did they help you answer the question?*

A student's performance on an item was determined once a student had provided an answer to the question during the concurrent portion of the interview. Some students would change their answers while being asked the 3 retrospective questions; however, their score on the 12 items had already been determined. All interviews were audio recorded and transcribed to maintain the accuracy of verbal reports. Appendix G contains the think-aloud interview instructions.

Data Analysis

The 16 protocols were transcribed by the researcher and a graduate research assistant. The initial practice question was not analyzed and therefore will not be presented in the results section. The concurrent portion of all of the student protocols was reviewed to determine if any systematic themes, defined as general cognitive processes, emerged. Themes were identified if they highlighted student strategy use, cognitive processing, and performance differences. After completing this initial examination of the concurrent think-aloud protocols, a number of themes were identified and named according to the strategy and/or cognitive process they elicited. These included (1) reading – rereads or comments on reading; (2) process of elimination with rationale; (3)

visualization; (4) school experience/knowledge; (5) life experience/knowledge; (6) feature – physical, graphic, or font; (7) attention to a word/concept/phrase; (8) indicates something was helpful; and (9) indicates something was confusing. The first seven strategies/cognitive processes were also identified for retrospective question one (*now tell me all that you can remember about how you solved this question*). Retrospective questions two and three resulted in simple yes/no responses followed, in some cases, by an explanation of what made the item confusing or helpful. The reported helpful and/or confusing features mapped directly on to the first seven identified themes. Based on these identified themes, a coding scheme was developed for analyzing the 16 protocols (see Appendix H). The researcher and research assistant employed the coding scheme with four student protocols to determine the utility of the coding system and to establish inter-rater agreement. Inter-rater agreement ranged from 82 to 84% and after discussion of disagreements, inter-rater agreement rose to 100% agreement. After all 16 protocols were coded utilizing the coding scheme, it was determined that the first seven themes could be collapsed under four general themes of general cognitive processes because they naturally fit together in terms of the student verbalizations. The four themes were: (1) comprehension (verbal and visual); (2) visualization; (3) background knowledge/experience (school or life); and (4) strategy use.

The comprehension theme included two sub-components, verbal and visual, that described student cognitive processing. The content of a student's verbal report was identified as "verbal comprehension" when they concurrently or retrospectively mentioned reading the item a number of times for clarification and/or paid additional attention to a particular word, concept or phrase when answering the question. For example, a female student (#14) repeated a word multiple times when concurrently reporting, "physical characteristics, what were, well we're learning about that right now. We're learning about physical and chemical characteristics…" The contents of student's verbal reports were identified with the code of "visual comprehension" when they mentioned the use of a physical feature, graphic, or font within the item to comprehend the item. For example, a male student (#13) stated "it is good that they italicized the least part because someone who was just reading it quickly might think they meant most so they give a completely different answer."

The second theme, visualization, was used to code students' verbal reports when they mentioned the use of visualization when solving an item. For example, on a question that had the students describe objects they might find around a pond, a male student #6 stated "I just thought what do ponds look like around the edges around their shores. I tried to visualize what I'd seen the last time I'd been around a pond."

When students concurrently or retrospectively reported the use of background knowledge and/or experiences in science, which could include school or life experiences, they received a code on the third theme. To illustrate this, on a question examining how a rock formation changed, a male student #5 said "I basically thought of what I learned in science and I thought that glaciers and earthquakes from my past knowledge that they don't happen much especially in a certain place neither does acid rain."

Strategy use, which appeared to be the most prevalent theme among the protocols, was coded when a student verbalized the use of process of elimination. To illustrate this, while student 8 was solving the first item he said "size of windows wouldn't really matter. Colour doesn't effect unless it is black. Number of layers could be possible. Type of material, I think it would be B." This theme is consistent with what the test-wiseness literature would describe as test-wise element 1D1 – eliminate options known to be incorrect (Millman, Bishop, & Ebel, 1965; Rogers & Bateson, 1991). However, for the purposes of this study the theme name of strategy use will be used throughout the results and discussion sections.

It should be noted that the think-aloud sample is small and is not representative of the large national sample that completed the SAIP science assessment. For these reasons, the results may offer an additional data to support student cognitive processing similarities and differences on these test items; however, they may not generalize to the national student sample. This limitation with the present study will be illustrated further in the discussion section.

Results

Results from the think-aloud verbal reports were used to further understand the nature of the dimensions found to underlie the SAIP science assessment and the different cognitive processes that male and female students employ while solving science test questions. For each item, the corresponding factor representation (ES, SR, CK, and RK), item content, difficulty level, and the gender that the item favoured are presented (see also Table 6). This is followed by a discussion of the student think-aloud protocol results. The concurrent results, followed by the results for retrospective questions 1, 2, and 3 will be presented for each of the 12 DIF items. The text for four of the items selected for the interviews cannot be presented because of the need for test security; these items may be included in a future PCAP assessment of science. However, the remaining eight items selected for the interviews have been publicly released (P. Brochu, personal communication, February 5, 2007). For the four withheld items, a general description will be provided to give the reader a flavour for the item followed by the results. See Appendix I for the think-aloud data frequencies by each theme and item.

Item one. This item was identified as a moderately difficult (difficulty level 3) item, and measures content related to knowledge and concepts of science – physics.

> To save energy, John's family decides to replace the windows in their home.
>
> **Which factor has the *least* effect on heat loss through windows?**
>
> A. Size of the windows
> B. Colour of the window frames
> C. Number of layers of glass
> D. Type of material separating the glass layers
>
> **Correct answer:** B

This item was identified as possessing C level DIF and favouring females on both the AB and AC tests. Furthermore, this item loaded on the reasoning with knowledge factor. Based on the verbal report data of the 16 participants, 6 of the 7 female participants answered the item correctly (85.7%) while 5 of the 9 male participants answered the item correctly (55.5%).

The concurrent portion of the interview revealed that a visual/verbal cue (the word *least)* aided females in reaching a correct response. Four of the male students skipped over the word *least* while reading, and based on their verbal reports it appeared as though they inserted the word *greatest* when determining their answers. Six of the male and 6 of the female students reported using the process of elimination strategy to aid them in generating a correct response.

When male students were asked to report all that they could remember about how they solved the item (retrospective #1), they introduced additional information that was not captured in their concurrent report. For example, 3 of the male students reported using visualization, 3 reported background knowledge/experience, and 3 reported process of elimination in determining their answers. On the other hand, only 2 females reported using additional processes that were inconsistent with their concurrent think-aloud protocols. On the final two retrospective questions (#2 – did you find any parts of this question confusing? and #3 – did you find any parts of this question helpful?), 4 of the 9 male students reported the word *least* as confusing. Appearing to contradict themselves, these same males in addition to the other 5 males identified *least* as helpful in the answering the question. Although the boys could identify *least* retrospectively as a helpful feature in answering the item, they were not all successful in answering the item correctly.

Item two. Item two cannot be displayed for test security reasons. This item is a multiple-choice item of moderate difficulty (level 3), and is referenced to the content area of knowledge and concepts of science – earth. This item was identified as favouring males and possessing C level DIF on the AB test and moderate DIF on the AC test. This item belonged to the scientific reasoning factor. Seven of the 9 male participants answered the item correctly (77.7%) and 5 of the 7 females answered it correctly (71.4%). It should be noted that this item includes a rather large graphic that 4 male and 2 female students concurrently reported using to solve the question correctly.

Within the concurrent part of the interview, 8 males reported the process of elimination with a rationale, and the use of the graphic in solving the problem. For example, a male student reported "I think wind and water cuz um wind and water like glaciers and earth quakes might not affect this area and acid rain might not go in this area either because there is trees." In addition to also utilizing process of elimination with rationale to solve the question, females reported using knowledge/experiences from science class to solve the question. For example, a female student commented:

> I think it's last year in science that we learned about this thing and it said there was a question like this and it was like a few ways of making rock moves some parts of it. And there was wind and water so that's possible but it's kind of high off the ground so I don't really think that's going to work. Glaciers and earthquake that kind of I don't know about that one. Acid rain that's kind of reasonable, and changes in temperature won't really help shape the rocks. So I think its wind and water because the particle things kind of like hit the rocks or something.

When male students were reiterating how they solved the item (retrospective #1), they again reported process of elimination and the use of the visual graphic. However, this time 8 males reported that what they learned in their science class was a contributing factor to answering the item correctly. Three males and 4 females reported that some of the response options were confusing. Three males and 3 females commented on attributes of the item such as the graphics, the words glacier and acid rain, and the item stem as confusing. On the other hand, 8 male and 6 female students identified that they found a part of the question helpful in answering the question; they identified that the graphic, the word *most likely* in the item stem, and the use of process of elimination were helpful.

Item three. This item was identified by SAIP as measuring knowledge and concepts of science – chemistry at a difficulty level of 2.

> **During their studies of the pond, the students walk along the shoreline and collect several objects.**
>
> **What is one physical characteristic, other than size or shape, that the students could use to classify the objects they find along the shore?**
>
> **Correct answer (any of the following would be considered correct): colour, density, hardness, living, non-living, magnetism, texture, floating, non-floating, and movement.**

This item possessed C level DIF on the AB test in favour of females and loaded on the scientific reasoning factor. The female students that participated in the think-aloud portion of the study outperformed the males. Six of the 7 females answered the item correctly (85.7%) and 7 of the 9 males answered it correctly (77.7%).

Two male and 3 female students concurrently reported the use of visualization when they answered this item. Four of these 5 students answered item 3 correctly. Furthermore, 3 males reported background experience related to visiting and walking around a pond, which appeared to aid them in their responses to the item.

The male and female student responses to retrospective #1 (tell me everything that you remember about how you solved this item) were relatively consistent with the concurrent reports. The students did not supply any new or contradictory information about how they solved the item; however, more students reported the use of visualization (7 males and 5 females). The confusing feature that was consistently mentioned by students was the background information that was presented prior to the question. Of the students that suggested something was helpful, 9 males and 4 females, all mentioned *other than size or shape* was a helpful feature.

Item four. This item measured knowledge and concepts of science – earth, and had a difficulty level of 3.

> Michelle knows that light reflected from the Moon's surface reaches Earth in about one second. She also knows that light from Alpha Centauri, the star nearest our solar system, takes about five years to reach Earth.
>
> **About how long does it take for light to travel from the Sun to Earth?**
>
> A. 1 second
> B. 8 minutes
> C. 5 years
> D. 10 years
>
> **Correct answer:** B

This item was found to possess C level DIF in favour of males on the AC test and B level DIF on the AB test. This item loaded on the everyday or elementary science factor. The male sample outperformed the female sample on this item, with 7 of the 9 (77.7%) males providing a correct answer and only 2 of the 7 (28.5%) females providing a correct answer.

During the concurrent reporting, 3 male students and 1 female student spent a lot of time reading and rereading the question and response options. In addition, 6 male and 4 female students attempted to use the process of elimination while rationalizing the response options to determine the correct response. Only one male student reported the use of background science knowledge while answering this question.

When the students were asked to report what they remembered about how they solved the question (retrospective #1), 3 male students reported using visualization, 5 males reported the use of background knowledge from science class, and 2 males reported the use of process of elimination. Four female students reported they used the process of elimination as their strategy for solving the item, and 1 female student reported she used her background knowledge in science. Five of the 9 male students and 4 of the female students reported that the words *Alpha Centauri* and the lack of information about the sun confused them when answering the question (retrospective #2). On the other hand, 8 males and 6 females reported that the information about *Alpha Centauri, the star nearest our solar system, takes about five years to reach earth* was helpful when answering the question (retrospective #3). This means that at least 8 of the male students and 4 of the female students contradicted themselves by saying that the same words were both helpful and confusing.

Item five. Item 5 cannot be presented for test security reasons. This item is a constructed response item that measures the nature of science and is considered to be at the easiest difficulty level (difficulty level 1). This item possessed C level DIF in favour of females and belonged to the everyday or elementary science factor. Males and females performed equally well, with the entire sample (9 males and 7 females) receiving perfect scores.

Ten of the 16 students did not think-aloud concurrently while solving this item and provided their answer immediately after reading the item. For those students that did not respond immediately 2 males and 2 females reread the item more than once prior to providing their answer.

When reiterating how they solved the item (retrospective #1), 3 males and 1 female reported the use of visualization and 3 males and 1 female reported the use of background knowledge and life experience. Of the 4 male students that reported something confusing in the item, specific words that they were unsure of the meaning were identified. Of the 6 males and 3 females that suggested something aided them in answering the question, specific information and words presented in the background information and item stem were reported.

Item six. This constructed response item was targeted to a difficulty level of 1 and was referenced to the content domain of science, technology, and society.

Brett and Bob cook their food over a campfire. They could have used a portable stove.

Name a possible fuel for each of the following.

	Campfire	**Portable Stove**
Fuel:		

Correct answers (any of the following would be considered correct responses): wood, paper, natural gas, gasoline, propane, butane, and coal.

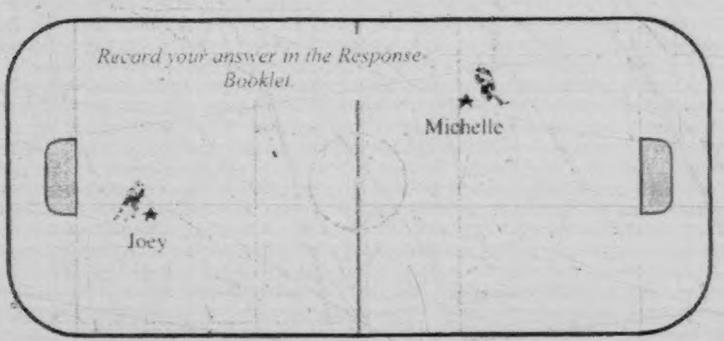
The item possessed C level DIF in favour of males and loaded on the everyday or elementary science factor. The 9 males provided a correct answer (100%) on this item while 5 of the female students answered the item correctly (71.4%).

Similar to the previous item and consistent with the response approach students used on easy items, 9 of the 16 students provided immediate responses without thinking aloud. For those students that provided concurrent reports, 4 male students and 2 female students reported they used their previous experience with camping.

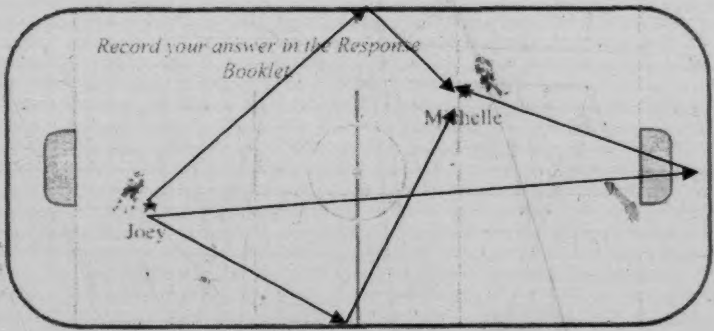
For retrospective question #1, 6 of the 9 male students reported knowing about fuels from science class and campfires from previous experiences. One male and 1 female student reported that they used visualization. For example, student #2 stated “I can make a diagram in my head about a campfire and sort of visualize little bits of information about what it needs and what it is using. With a portable stove I think about the parts that you need to run a portable stove and classify which ones are used as fuel.” One female student reported knowledge of her gas stove at home while a different female student reported never camping in her life. Four of the 7 female students reported that this question was confusing and indicated the words *portable stove* and *fuel* were confusing. Although 4 of the male students suggested that something was helpful in answering the item, only 2 male students could articulate that it was a result of the format of the question (constructed response) and the word *portable stove* that was helpful.

Item seven. This item measured knowledge and concepts of science – physics and was targeted at a moderate difficulty level (level 3).

> Every winter after school, Richard plays a game of pick-up hockey at the local arena with his friends Joey and Michelle. Joey shoots the puck off the boards, and it goes to Michelle.
>
> **Draw the lines to show the path of the puck as it goes from Joey to the boards and then to Michelle.**


Record your answer in the Response Booklet.
Michelle
Joey

Correct answer:


Record your answer in the Response Booklet.
Michelle
Joey

Item 7 was found to systematically favour males on both the AB and AC tests (C level DIF). Moreover, the item loaded on the reasoning with knowledge factor for the AB test.

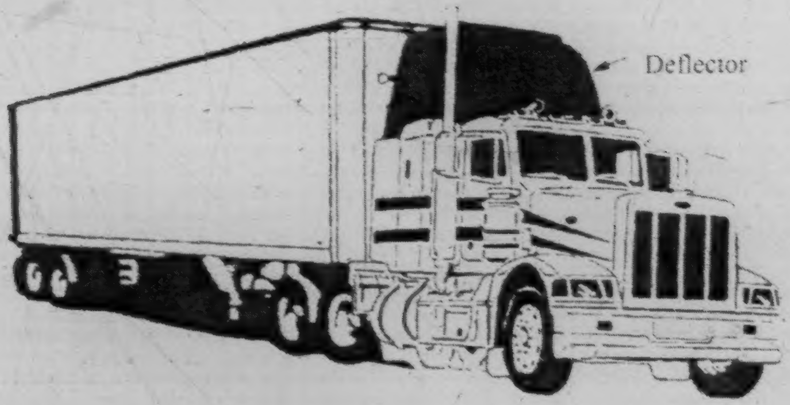
All of males correctly answered this item whereas only 5 (71.4%) of the females responded correctly.

During the concurrent portion of the interview, 4 male students reported that the picture was helpful in answering the question, 2 males reported visualization, and 5 males reported the use of background knowledge and experience to solve the item. The male students reported their knowledge of mirrors and angles of incidence and reflection from science class, and their life experience playing billiards and hockey when answering this item. Two female students reported their life experience with indoor soccer and billiards while solving the item.

When recalling how they solved the items (retrospective #1), 7 males reiterated their use of background science knowledge while 2 males reported background life experiences (for example, their experience with billiards and hockey). Four of the female students recalled that they used school or life experiences. When asked if they found any parts of the question confusing (retrospective #2), 3 male students reported that they found the picture confusing in solving the question as one student illustrated "it doesn't really say which boards to use. Do you use the one on the right or the one on the left?" Two female students identified that the additional information provided prior to the question was confusing and unnecessary. Seven males and 5 females identified features that helped them in answering the question. The features that were most consistently reported as most helpful were the picture and how it was labeled, and the statement "off the boards" in the background information.

Item eight. This item measured concepts of science- earth and was identified at difficulty level 2.

When Joey and Michelle leave, they see a truck at the back of the hockey area. It has a large deflector over the cab.



What is the deflector used for?

A. To reflect the Sun's rays
B. To reduce the noise level
C. To advertise the company logo
D. To direct the air flow above the trailer

Correct answer: D

This item possessed C level DIF in favour of males and belonged to the everyday or elementary science factor. Contradictory to the DIF results, the female student's in the interview sample outperformed males on this item with 6 of the 7 (85.7%) female participants getting it correct, and only 5 of the 9 (55.5%) males getting it correct.

The concurrent portion of the interviews indicated that the majority of male and female students utilized the process of elimination with rationalization (6 male and 6 female students). To illustrate this with an example, a female student reported:

> For the part where it says to advertise the company logo I don't really think it's that one because they could just do it on the side of the truck and it wouldn't have to be called a deflector. To direct the airflow above the trailer if it's that one but it doesn't really seem likely. Actually maybe that one because I don't really think any of the other ones it is used for.

In addition, 3 male students concurrently reported that the word *deflector* and the picture were helpful while solving the item.

When recalling how they solved the item (retrospective #1) 5 female students reported diverse words and identifying features in the item that contributed to their thought processes (for example, *direct air flow, it's black, see the shape,* and *deflector).* Although not reported in the concurrent portion of the interview, 2 male and 2 female students reported that they used visualization, and 3 male and 1 female student reported that they used background knowledge and experiences. When asked if any parts of the item were confusing (retrospective #2), 3 male and 3 female students reported that the words *deflector, reduce noise,* and *direct airflow* confused them. On the other hand, 8 males and 5 females reported that the picture and the arrow pointing to the deflector were helpful components.

Item nine. Item 9 cannot be presented because of test security reasons. This item is a multiple choice item targeted at difficulty level 3 and measures the content domain of nature of science. This item was found to systematically favour female students on both the AB and AC tests as evidenced by B level DIF. This item loaded on the scientific reasoning factor. Six of the 9 (66.6%) males responded correctly and 5 of the 7 (71.4%) females responded correctly.

Seven males and 4 females engaged in the process of elimination while providing a rationale for their responses when they answered this item. Three male students had to reread the question more than once as they appeared not to comprehend what it was they were supposed to do. One male student mentioned that his father is a medical doctor and used his knowledge about science and medicine to answer the question.

Somewhat consistent with their concurrent reports, 7 male and 3 female students reported that they used the process of elimination with rationale to solve the item when the students were asked to recall how they solved the item. In addition, 4 males and 1 female reported that they answered this question based on their knowledge about medicine from visiting the doctor, a friend that takes insulin, or having parents that were doctors. Finally, 1 male and 1 female student stated that they did not know the meaning of *control group*. When asked if any parts of the question were confusing (retrospective #2), 5 male and 4 female students reported that the *control group* confused them. Six male students reported two features of the item that they found helpful when they answered the question; the word *not necessarily* and the multiple choice options that were available. These male students reported that the multiple choice options helped them narrow down the choices to come up with the correct answer. Of the 3 females that reported that something was helpful in the item, 2 reported the fact that there were multiple choice options that were worded in such a way that helped them answer the question.

Item ten. Item 10 cannot be presented due to test security reasons. This item is a multiple choice item that was referenced to the content domain of nature of science, and is a more difficult item (level 4). This item possessed C level DIF in favour of females. It loaded on the chemistry knowledge factor. The females outperformed the males with 5 out of 7 (71.4%) correct responses for females and 5 out of 9 (55.5%) correct responses for males.

The concurrent reports revealed that 4 males and 4 females reread the item more than once for clarification. Furthermore, 5 males and 4 females reported they used the process of elimination with a rationale to select their answers. For example, a female student reported

> 'A' doesn't make sense, to enhance her observations and modify the hypotheses.... O.k. theories would be for accurate explanations for conclusions. It's either B or D. To enhance her observations and modify. Just doing a test would enhance her observations. More reliable hypotheses, I think it's D.

When asked to recall how they solved the item, none of the students who reread the item for clarification reported this as a strategy. However, they did consistently report the use of the process of elimination to arrive at their answers. Although 6 of the 9 males and 2 of the 7 females reported that something was confusing in the item, they were unable to articulate what it was. On the other hand, 5 males and 5 females consistently identified words (such as *she constantly tests, sets aside scientific theories,* and *explanations for* conclusion) in the item stem that helped them.

Item eleven. Item 11 measured the content domain of science, technology, and society and is a more difficult item (level 4).

Gene Therapy
A 30-year-old woman was suffering from a rare and deadly genetic disorder. She lacked a gene that enabled her liver to remove low-density lipoproteins, often called "bad" cholesterol, from her blood.

In an experimental treatment, a portion of the woman's liver was surgically removed and researchers inserted properly working genes into its cells. The portion of the liver was then transplanted back into her body.

What is the most direct use of this type of gene therapy?
A. To improve human physical fitness
B. To treat viral diseases such as the common cold
C. To improve beef as a low cholesterol food source
D. To synthesize chemical substances normally produced by the body

Correct answer: D

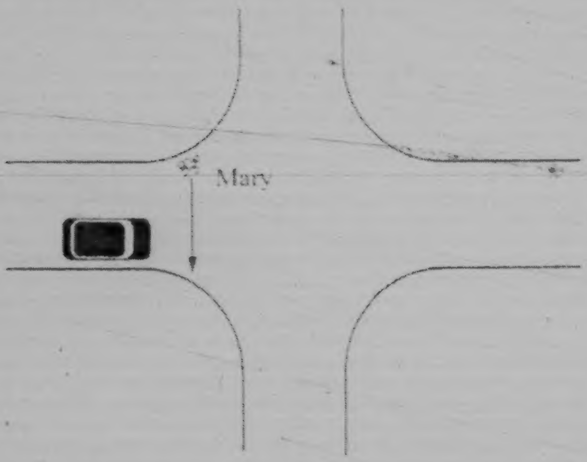
This item possessed C level DIF in favour of females and belonged to the scientific reasoning factor for the AC test. Five of the 7 (71.4%) females answered the item correctly while 6 of the 9 (67%) males answered the item correctly.

When students were asked to think out loud as they solved this item, 3 male students and 1 female student responded immediately to this item, while the remaining males and females reported they used the process of elimination.

Although it was not mentioned during the concurrent reporting, when the students were asked to recall how they answered the question, 2 male and 2 female students identified the use of visualization as a problem solving strategy. For example, a male student retrospectively reported "I remember picturing a person actually going through this and I am thinking it was a genetic disorder and not a viral disorder and that was one of my top choices and I ended up crossing it out." Of the 7 males and 5 females who reported they used process of elimination during their concurrent reports, only 3 males and 3 females recalled using this strategy. One male student reported remembering learning about this in science class and that his "grandmother has liver problems." When asked if any parts of the question confused them (retrospective #2), 2 male and 3 female students said yes, and they indicated that the words *synthesize* and *low density lipoproteins* were confusing. On the other hand, 4 male and 4 female students identified that utilizing words in the item stem and background story (*most direct use* and *researchers inserted properly)* helped them.

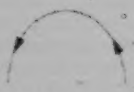
Item twelve. The final item in the think-aloud data set measured knowledge and concepts in science – physics and was considered the most difficult item (level 5). This item possessed C level DIF in favour of males and loaded on the scientific reasoning factor for the AC-13 test. Males out performed females with 5 of the 9 (55.5%) males getting the item correct and 3 of the 7 (42.8%) females getting the item correct.

Rita's car stops at an intersection. Mary walks straight across the street in front of the car at a constant speed, as indicated below. While walking, Mary throws a ball vertically in the air.

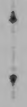


Which trajectory of the ball does Rita see from inside the car?

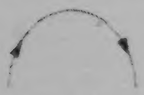
A.



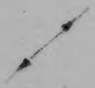
B.



C.



D.



Correct answer: C

The students reported a number of strategies while solving this item. Four male and 5 female students commented on the word *vertically* as well as the picture when solving the item. Six male and 4 female students reported using visualization, such as imagining themselves in the car and watching Mary. For example, a male student concurrently reported:

> In my head I am thinking that the question is asking from the person inside the car how did they see the ball bouncing from one side to the other side. So I am remembering myself sitting in the car with my parents and seeing people walking across and I am just imagining a ball going across the street and from everything that I see from inside the car I don't think anything changes. So I think the answer is B because I don't think that from inside the car there is no big change in trajectory from one place.

Five of the 9 male students used the process of elimination with rationalization and the arrows in the response options while they answered the question.

Although 9 males reported the use of process of elimination, it was interesting that only one of the male students described using it when they were asked to recall how they solved the question (retrospective #1). Male and female students were, however, consistent in reporting the use of visualization when asked to remember how they solved the item. Three male students identified that the pictures and the curvature of the windshield were confusing while 3 female students identified the word *trajectory* as confusing. In contrast, 5 male and 5 female students identified that the diagram and response option pictures were helpful in solving the item.

Summary of Verbal Report Data

The think-aloud verbal reports provided an additional data source with which to highlight the differences in male and female strategy use and cognitive processing when solving the 12 DIF items. By having male and female students think out loud as they solved the 12 items identified as statistically significant for DIF, it became clearer in identifying why some items would have the potential to create gender performance differences. The first DIF item was found to statistically favour female students on the SAIP and the verbal report study also revealed that more female students answered the item correctly. The majority of female students (6 of 7) and 5 of the 9 male students that concurrently reported attention to detail in reading and a focus specifically on the word *least* answered the item correctly. Furthermore, students reported that the word *least* was helpful in generating a correct response suggesting that this could be a potential source of DIF.

The second DIF item could not be displayed in the text because the item had not been publicly released. This item was found to statistically favour male students; however, the verbal report data revealed that there were no gender differences in overall performance on this item. This item possessed a pictorial graphic that more male students concurrently reported using when solving the item. On the other hand, the female students that answered the item correctly reported the use of their background knowledge in science class while solving the item. Although more male than female students used the pictorial graphic to answer the item and this could be considered irrelevant to the construct being measured, just as many female students answer the item correctly relying on their knowledge in science. Based on the verbal reports, it is not clear what the potential source(s) of DIF was.

The third DIF item was found to statistically favour female students; however, the verbal reports resulted in as many male and female students answering the item correctly. Although, both male and female students concurrently reported the use of visualization and background life experiences when solving this item it was not sufficient evidence to make any claims about the source of the DIF in this item.

The fourth DIF item was found to statistically favour male students and the verbal report data supported this finding with 7 of the 9 male students and only 2 of the 7 female students responding correctly. On average, students took more time trying to solve this item, they read and re-read the question and response options numerous times. While male and female students reported that certain words and phrases were helpful in solving the item, there was very little information to support a conclusion about the reason this item was found to statistically favour males.

The fifth DIF item was found to favour female students and the verbal reports revealed that all male and female students answered the item correctly. This item required students to read an excerpt about a scientist conducting an experiment with sun screen and the question was geared at understanding how best to protect the participants in her study. This item did not lend itself to concurrent reporting, possibly because it was at the easiest difficulty level. Given that all students answered the item correctly and the concurrent and retrospective reporting was sparse, no conclusions could be generated about the source of DIF with this particular sample.

The sixth DIF item in the verbal report portion of the study was found to statistically favour males and the verbal report data revealed that all male students answered the item correctly while only 5 of the 7 female students did. Similar to item 5, more than half of the students answered this question immediately without providing any concurrent data. Male students concurrently and retrospectively reported using background knowledge from both science class and their life experiences. Furthermore, male students reported that the words *portable stove* were helpful in answering the item while a number of female students reported that these words were confusing. Based on the verbal reports, one potential source of DIF could be the differences in knowledge of specific terminology that male and female students possess.

The seventh DIF item was found to statistically favour male students and again all male students received a correct response. The male students concurrently reported the use of the diagram of the hockey rink, visualization, background knowledge in science, and their life experiences playing billiards and sports as helpful aspects that aided them in answering it correctly. Of the 5 females that answered this item correctly, 2 female students reported their experiences with indoor soccer and billiards as helpful while answering the item. Based on the verbal report data, it could be hypothesized that while knowledge of sports is irrelevant to the construct being measured it did help male students respond correctly, and therefore could be the source of the advantage.

The eighth DIF item was found to statistically favour males; however, the verbal reports revealed that the female students outperformed the male students on this item. The majority of students, both male and female, reported the use of process of elimination to answer the item correctly. Three of the male students concurrently reported that the word *deflector* was helpful in determining a correct response. On the other hand, 3 male and 3 female students reported that the word *deflector,* in addition to the words *reduce noise,* and *direct airflow,* were confusing. This item relies heavily on the use of a pictorial graphic to answer the item correctly, which may not be relevant to the construct being measured and in turn could create a potential form of bias for those students that could not use the picture to their advantage. Although a number of students verbally reported that the picture of the truck and the arrow pointing to the deflector were helpful it was not clear whether this was the source of DIF.

The ninth DIF item was found to favour females and the verbal report data revealed that 6 of the 9 males and 5 of the 7 females answered the item correctly. This item could not be displayed in the text as it has not been publicly released. Male and female students concurrently and retrospectively reported the use of process of elimination to answer this item. A number of students focused on the word *control group* and reported that they either did not know the meaning of the word or it confused them when trying to answer the question. Given that the word *control group* is present in the correct answer could potentially advantage those students with knowledge of the meaning of this word. Although not entirely supported through the verbal report data collected, the meaning of this word could be a potential source of DIF and would not be biased but rather impact – a true ability difference in knowledge.

The tenth DIF item was also found to favour females and could not be presented in the text because it has not been publicly released. The verbal reports revealed that more female than male students answered this item correctly (5 female and 5 male). This item was language rich in that it possessed a lot verbal information for each response option. This claim was supported in throughout the concurrent reporting in that a number of students read and re-read the question and response options a number of times. Furthermore, a number of students reported that words and phrases such as *she constantly tests, sets aside scientific theories,* and *explanations for conclusions* were helpful in answering the question. With that said, there were no consistent findings within the verbal report data collected from the 16 students that could identify why females statistically were found to perform better on this item.

The eleventh DIF item was also found to statistically favour females and was language rich. The verbal report data indicated that 6 of the 9 male and 5 of the 7 female students answered the item correctly. Four students (3 males and 1 female) responded immediately to this question even though it was not considered an easy item. This item produced similar results as item 10, in that, there was limited verbal report data from the 16 students to generate hypotheses about why the item was found to statistically favour females in the larger SAIP sample.

The final DIF item used in the verbal report portion of the study was found to statistically favour males. The concurrent verbal reports revealed that both male and female students did not perform very well on this item with 5 of the 9 males and 3 of the 7 females getting the item correct. This item was similar to the DIF item 8 in that the pictorial diagram was essential in generating a correct response. Furthermore, a number of students reported the use of visualization and the word *vertically* in the item stem as helpful while solving the item. A student's ability to use a pictorial diagram to answer this question correctly could be a source of DIF only if it is found that the diagram is not relevant to the construct being measured. The use of visualization could also be a source of true ability difference, however, given the limited verbal report data and restricted sample of students in the study this is merely a hypothesis and not conclusive.

Chapter Five: Discussion and Conclusions

Chapter 5 is divided into three sections. The first section provides an overview of the purpose and research questions that were addressed in this study. The second section contains a summary of the methods used and findings that were obtained. In the third section, the utility of the three stage approach to understanding the underlying constructs that are measured by the SAIP Science assessment in the context of the existing research is discussed. Lastly, a discussion of the limitations of the study and implications for future practice and research will be provided.

Purpose and Research Questions

Large-scale assessments are used for a variety of reasons provincially, nationally, and internationally. Some of these purposes include the provision of information about how a student performs relative to their peers and/or how a group of students compares to another group of students at a national and international level. Given the possible implications of these comparisons, it is imperative that the inferences generated from test scores are in fact a valid representation of student performance. The purpose of this study was to investigate the utility of a multi-method approach for collecting validity evidence about the underlying knowledge and skills measured by the 2004 version of the School Achievement Indicators Program (SAIP) Science Assessment. The three approaches included analysis of dimensionality, differential item functioning (DIF), and think-aloud interviews. These approaches were pursued in an attempt to investigate and explain the construct of science achievement in order to provide validity evidence for the inferences that are generated from test scores. The specific research questions addressed were:

1. Does the 4-factor model previously found by Hamilton et al. (1995) for the grade 8 sample explain the SAIP 2004 data? If not, what is the dimensional structure of the 2004 version of the SAIP Science assessment?
2. Do the performances of male and female students on the SAIP items systematically differ? If dimensions are found, are performance differences systematically captured in the dimensions?
3. If performance differences are found, can interview data of male and female grade 8 students aid in the generation of hypotheses about the underlying knowledge and skills that are measured by this test?

Method and Summary of Findings

The first research question was addressed by investigating whether the grade 8 English sample on the SAIP Science Assessment fit the 4-factor model put forth by Hamilton et al. (1995) using confirmatory factor analysis. The second research question was answered by analyzing the two tests (AB and AC) for differential item functioning (DIF). The final question was addressed by collecting think-aloud data of students reporting their thought processes as they solved a selection of items from the AB and AC tests. Given the sequential nature of the study, each method and findings for that method are presented together for the corresponding question.

Question 1: Can the 4-factor model previously found by Hamilton et al. (1995) for the grade 8 sample explain the SAIP 2004 data? If not, what is the dimensional structure of the 2004 version of the SAIP Science assessment?

Question 1 was addressed by conducting a confirmatory factor analysis using LISREL 8.53 (Jöreskog, & Sörbom, 2002) of the 4-factor model found in Hamilton et al.'s (1995) study. The 129 items that made up the AB and AC tests were coded according to Hamilton et al. (1995) and Nussbaum et al.'s (1997) factor descriptions and the ratings derived by Leighton et al. (2007). The values for the four fit indices used – RSMEA, RMR, AGFI, and the chi-square statistic – revealed that the 4-factor model fit the data well for both the AB and AC tests. This finding is basically consistent with previous findings of dimensionality in science achievement (Ayala et al., 2002; Hamilton et al., 1995; Leighton et al., 2007; Nussbaum et al., 1997). While the SAIP 2004 data could be explained by the 4-factors retained in Hamilton et al.'s (1995) study, the distribution of items across the 4-factors in the present study does not entirely reflect the distribution of items found by Hamilton et al. Of the 25 items examined on the NELS: 88, 12 items loaded on the everyday or elementary science factor (ES), 5 items loaded on the scientific reasoning factor (SR), 4 items loaded on the chemistry knowledge factor (CK), and 4 items loaded on the reasoning with knowledge factor (RK) (Hamilton et al, 1995). In the present study, the majority of items were found to load on the ES and SR factors for the AB test. Conversely, for the AC-13 test the majority of items loaded on the SR and CK factors. This result could reflect the differences in difficulty of the AB and AC tests. The AB test is assumed to measure lower level science achievement abilities (CMEC, 2007), which would be consistent with the ES factor. On the other hand, the AC test measures higher level science achievement abilities, which would be reflective of the SR and CK factors. The findings in this study are also consistent with research into the dimensional structure of a previous SAIP administration (1999), which found that the 4-factor model also fit both the AB and AC tests (Leighton et al., 2007). The use of Hamilton et al.'s 4-factor model increases interpretability of the factors in light of the knowledge and skills being measured, which was argued by Gorsuch (1983) as one of the most important things to consider in factor analysis.

Question 2: Do the performances of male and female students on the SAIP items systematically differ? If dimensions are found, are performance differences systematically captured in the dimensions?

Previous research into gender differences in science has indicated that males and females systematically differ in their use of reasoning and response strategies (Beller & Gafni, 1996; Halpern, 1997; Hamilton, 1998; Hedges & Nowell, 1995, Linn & Peterson, 1985). In this study, SIBTEST (Shealy & Stout, 1993a) was employed to determine if the performance of male and female students differed systematically on the SAIP items. The SIBTEST results indicated that 24 of the 83 (29%) AB test items and 16 of the 77 (21%) AC test items displayed DIF. Using the guidelines put forth by Roussos and Stout (1996), 13 items possessed large or C level DIF (six favoured males and seven favoured females) and 11 items possessed moderate or B level DIF (eight favoured males and three favoured females) on the AB test. Seven items possessed large DIF (three favoured males and four favoured females) and nine items possessed moderate DIF (four favoured males and five favoured females) on the AC test.

Of the ten items that favoured females on the AB test (B or C-level DIF), 6 items corresponded to the scientific reasoning (SR) factor, 3 items to the everyday or elementary science (ES) factor, and 1 item to the reasoning with knowledge (RK). Two of the items that favoured females appeared on the surface to measure scientific reasoning that would be applied to formal science concepts (SR), but, based on the think-aloud reports, these items actually required attention to detail in reading. Although this result does not translate across all SR items found to favour females, it does support at least one hypothesis suggesting that females outperform males on items that require careful reading and attention to detail (Klein et al., 1997; Stumpf & Stanley, 1996). Of the 14 items that favoured males, 8 items represented the SR factor, 5 items represented the ES factor, and 1 item represented the RK factor. There were no DIF items found on the AB test that represented the chemistry knowledge factor (CK), which is not surprising given that only 9 items belonged to this factor. The items that favoured males tended to be related to earth and space sciences, stereotypical male related activities (camping, hockey, and trucks), and numerical operations (measurement and calculations). These items correspond to the types of items found in previous studies that revealed large male advantages on items related to physical science and earth and space science (Beller & Gafni, 1996). This is also somewhat consistent with Hamilton's (1998) findings that suggested that males outperform females on TIMSS items related to the physical and earth sciences.

Of the 9 items that were found to favour females on the AC test, 8 items represented the scientific reasoning (SR) factor, while 1 item represented the reasoning with knowledge factor (RK). Of the 6 items that favoured males, 3 items represented the SR factor, 2 items represented the CK factor, and 2 items represented the ES factor. The gender difference studies that were conducted previously using the factor structure recovered from the NELS: 88 focused on the 3-factors found for the 10th and 12th grades where they attributed the gender differences to the spatial mechanical (SM) factor (Hamilton et al., 1995; Hamilton et al., 1997; Nussbaum et al., 1997). The SM factor is primarily made up of physical, earth, and space science items, which is consistent with the male advantage found in the present study. There were no differences found among the quantitative reasoning (QR) and basic knowledge and reasoning (BKR), which is not consistent with the present study considering that the BKR factor is primarily made up of ES and SR items. Previous research has focused on the degree to which item format impacts male and female performance (Hamilton, 1998; Resnick & Resnick, 1992), the DIF items found in this study did not support the earlier finding that there were performance differences due to item format.

Question 3: If performance differences are found, can interview data of male and female grade 8 students aid in the generation of hypotheses about the underlying knowledge and skills that are measured by this test?

There are few studies that have employed think-aloud techniques with educational tasks and more specifically within the domain of science (Baxter & Glaser, 1998; Ercikan et al. 2004; Hamilton et al. 1997; Leighton & Gokiert, 2005). Concurrent and retrospective verbal reports (Ericsson & Simon, 1993) were collected from a sample of nine male and seven female grade 8 students. The concurrent and retrospective think-aloud data of the examinees, while they solved 12 items that possessed large or moderate DIF, were used to determine the underlying knowledge and skills that are measured by the SAIP Science test and how male and females respond differently to test items using these skills. The data were examined to determine if these differences in responding deducted or contributed to overall performance. Four general cognitive processing themes were identified that could be used to explain male and female problem solving on the SAIP Science assessment. The four themes included (1) comprehension (verbal and visual); (2) visualization; (3) background knowledge/experience (school or life); and (4) strategy use.

The comprehension theme was present in ten of the think-aloud items. Five of the items required verbal comprehension because they possessed a heavier reading component and necessitated attention to words, concepts, and phrase details in order to comprehend what the question was. For example, within the item stem words such as "least", "not necessarily" and "most likely" were italicized to emphasize their importance in answering the question. Both male and female students answered the questions correctly when they focused their attention on italicized words and/or engaged in reading and re-reading. Males reported the use of a graphic or physical feature in the item when completing five of the items that were visually rich. Lastly, when students were asked to remember how they solved the items, they in essence used metacognitive strategies to think about how they solved the item and, based on the findings in this study, they would inevitably recognize the important pieces of information crucial to problem solving.

Students also reported strategies that could be categorized as visualization, the second theme, on three of the items, while in their retrospective reports the students identified the same three items and an additional five items in which they used visualization. Although both male and female students used visualization, the use of visualization was more prevalent among the male students, which is consistent with the comprehension of visual graphics of features discussed in the previous paragraph. Further, this finding is consistent with previous findings that have demonstrated gender differences on items that possess visualization requirements when problem solving (Hamilton, 1999).

The third theme involved the use of the students' background knowledge and experiences, which were reported as learned in their day-to-day lives, at school, and in the classroom. During concurrent reporting, very few (approximately 2 to 5) students mentioned actual science knowledge and experiences when answering the questions; they were more apt to describe life experiences such as camping and sports that contributed to their problem solving. Interestingly, when students were asked to retrospectively describe how they solved the items, they frequently reported information they were currently learning in science class, previously learned in school, and specific life experiences. The two sets of finding are consistent with the finding of previous research conducted by Leighton and Gokiert (2005); it was expected that students would report that their knowledge in science and experiences in science class would contribute to answering an item. This outcome suggests that some of the SAIP items may be dependent on science knowledge; however, at the time of responding, students may not articulate this. However, students could retrospectively articulate the type of content, knowledge, and skills they "thought" they used when not directly engaged in problem solving (i.e., concurrently reporting their thoughts).

The final theme involved strategy use. When describing how they determined item responses, both male and female students engaged in and described the use of the process of elimination – the act of discarding item options based on their knowledge and experiences. Furthermore, on more difficult items students concurrently reported the use of process of elimination using previous background knowledge and experiences to eliminate options known to be incorrect. For the male students that used process of elimination as a strategy, the majority provided a reason only for why they thought the alternative they chose as their final response was a reasonable answer. On the other hand, female students provided reasons for why they eliminated an item alternative as opposed to why they selected their final answer. Although the female students seemed confident in their decisions to reject an alternative as observed during the interview, they were more hesitant when selecting their answer. When determining the answers for constructed-response items, which do not allow for direct process of elimination of alternatives, females would go so far as to continue to provide probable alternatives, and reasons why they were not selecting a particular answer. Additionally, test taking strategies, such as the use of test wiseness were used throughout the items and contributed to improved performance (Rogers & Bateson, 1990). The use of test-wise skills and relevant knowledge in science, can help students to perform quite well when approaching moderately difficult SAIP Science items.

The think-aloud data also contributed to formulating hypotheses about sources of DIF, whether it was construct relevant or irrelevant, due to bias, or due to impact. Although some items possessed bias it, alone, did not always support the statistical findings. In other words, although an item may have been found to favour females statistically, during the think-aloud portion of the study both male and female students performed equally well on the item. For example, in the first think-aloud item the word *least* was italicized and created a qualifier that was critical to answering the item correctly. This item was found to statistically favour females and on average females were more apt to pay attention to the word *least* during the think-aloud interview. However, for those male students that were more attentive to details, they also answered the question correctly. Additionally, items 3, 8, and 9 also lent themselves to attention to detail in promoting higher performance. Although these results cannot be generalized beyond the sample of students in this study, it appeared as though males possessed less attention to detail in reading and would become confused or less interested in focusing on an item that possessed a lot of reading. Males tended to perform better than females on the items that required them to interpret or use a visual or diagram to answer the question. The items that lent themselves to visualization, such as items 3 and 12, resulted in performance differences. Visualization was a cognitive skill that students possessed and applied to items in order to answer the item correctly.

Limitations

The SAIP is a low stakes large-scale assessment with little to no consequences for poor performance, and validation studies such as these may be excessive for the types of inferences that are generated from low stakes tests of this nature. However, validation of low-stakes tests can help researchers explore the best sources of evidence to use in validation arguments. This would be more challenging to do with high-stakes tests because tests of this nature are more protected and less likely to be available for analysis. A delimitation of this study is that the 4-factor model outlined in Hamilton et al. (1995), which was chosen to determine if the SAIP could be explained by more than one dimension, is not the only model that could potentially fit the SAIP 2004 data. Additional models could further explain the knowledge and skills measured by this test.

Several limitations reduce the generalizability of the results of this study. First, in the think-aloud portion of the study, few items were selected (12 items) and although they are representative of all of the 4-factors, they may not have allowed for subtle cognitive differences that would be present in other items. Given this, the hypotheses generated from student reports may not generalize to all items on the test. Second, the gender differences that were drawn from the think-aloud interview, appeared to be very item specific and the findings would likely not inform the remaining DIF items found on the AB and AC tests. Furthermore, the items selected for the think-aloud portion of the study did not include non-DIF items. This does not allow for a direct comparison of male and female thought processing while solving items that could be considered biased from those that do not statistically represent a probability of one group performing better than another.

Finally, the sample of students that was selected to complete the think-aloud portion of the study was different from the original 2004 SAIP sample in at least two ways. The students selected for the think-aloud portion of the study came from a school that had a focus on metacognition and higher level thinking skills; therefore, they may have been primed to engage in verbal think-aloud techniques prior to being interviewed. Based on their classroom cumulative grades, these students would be considered high achieving. Consequently it is not known how low performing students may interpret and respond to the test items considered. However, it is also the case that low performing students may not be able to articulate their problem solving strategies. Finally, the sample size of the think-aloud study is too small to statistically determine whether the differences between males and females were significant and replicable.

Conclusions

Despite the limitations of the present study, the multifaceted approach produced interpretable and meaningful validity evidence about the knowledge and skills students apply on the SAIP Science Assessment. The dimensionality analysis confirmed that the

4-factor model found on the NELS: 88 science assessment could be used to explain the underlying knowledge and skills measured by the SAIP 2004 science test. That the dimensions found in this study could be substantively defined by 4-factors (ES, SR, CK, RK) more meaningful inferences about student performance could be generated from the SAIP data. The DIF analysis revealed that a number of items possessed moderate to large differences for male and female students on the SAIP Science assessment. The think-aloud verbal reports provided an additional data source with which to highlight the diverse knowledge, skills, and strategies male and female grade 8 students use when solving science test items. Even though dimensionality and DIF analyses and think-aloud verbal reports yield comprehensive forms of validity evidence, as was demonstrated in this study, they only begin to provide a basic understanding of the underlying construct(s) that are being measured by the SAIP Science assessment.

Implications for Practice and Future Research

The validity evidence that was collected in this study could be further examined in a number of ways that would enhance the findings and test development generally. First, if additional think-aloud data could be collected from a representative sample of AB and AC SAIP items that reflect the 4-factors, this would provide further substantive support for the selection of the 4-factor model to describe grade 8 science achievement on the SAIP Assessment. This type of research could lend further support to the claim that science is a multidimensional construct and therefore, it may be more informative to present students with subtest scores that would speak to their strengths and weaknesses in different domains of science.

In addition, it would be useful to have a more representative sample of students think-aloud on a larger number of DIF and non-DIF items to adequately capture the unique cognitive processes male and female students across the achievement spectrum use when problem solving on tests of science achievement. A complimentary technique that would augment the hypotheses generated about the construct relevancy of the dimensionality found within specific DIF items would involve engaging a group of experts in science to review the DIF items. This would provide further substantive evidence about the DIF that was found and how it relates to the overall dimensions found in the study. This would arm developers of large-scale science assessments with more knowledge about the potential biases inherent in science test items, and the cognitive differences that male and female students possess. With this knowledge, test developers would be in a better position to make evidence-based decisions about which items to retain, modify, or discard.

Being mindful of the types of items that are included in think-aloud studies may provide more evidence to support hypotheses about unique science domains, cognitive processing, and skill differences. It would be appealing to examine if the 4-factor model also fits the data collected from the first administration of the Pan-Canadian Assessment Program (PCAP) and other large-scale science assessments such as the TIMSS and PISA. Finally, it would be appealing to examine the utility of the three approaches described in this study on a high-stakes standardized achievement or cognitive assessment.

References

Alberta Education (2005). *Public information.* Retrieved February 18, 2005, from http://www. education.gov.ab.ca.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Ayala, C.C., Shavelson, R.J., Yin, Y., & Shultz, S.E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment: Issues and Practice, 8*(2), 101-121.

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., & Kelly, D.L. (1996). *Science achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88,* 365-377.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42* (1/2), 1-21.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models.* Newbury Park, CA: Sage Publications.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models.* Newbury Park: Sage.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *The Journal of the Learning Sciences, 6* (3), 271-315.

Childs, R.A., & Oppler, S.H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement, 60*(6), 939-955.

Chudowsky, N., & Pellegrino, J.W. (2003). Large-scale assessments that support learning: what will it take? *Theory into Practice, 42*(1), 75-83.

Cizek, G.J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

College Board (2005). *The new SAT 2005.* Retrieved April 14, 2005, from http://www.collegeboard.com/newsat/index.html.

Council Ministers of Education, Canada (2000). *Public report on science assessment: SAIP School Achievement Indicators Program 1999*. Retrieved August 12, 2004, from http://www.cmec.ca/saip/science2/science2.en.stm.

Council Ministers of Education, Canada (2004). *Public report on science assessment: The SAIP Science III 2004 Assessment.* Retrieved August 15, 2005, from http://www.cmec.ca/pcap/science3/indexe.stm.

Council Ministers of Education, Canada (2007). *The PCAP-13 2007 Assessment.* Retrieved May 23, 2007, from http://www.cmec.ca/pcap/2007/indexe.stm.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

DeMars, C.E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.

Downing, S.M., & Haladyna, T.M. (1996). A model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coup. *Educational Measurement: Issues & Practice, 15,* (1), 5-13.

Embretson, S.E. (1999). Cognitive psychology applied to testing. In F.T. Durso (Ed.), *Handbook of applied cognition* (pp.629-660). Chichester, England: John Wiley & Sons.

Enman, M., & Lupart, J. (2000). Talented female students' resistance to science: an exploratory study of post-secondary achievement motivation, persistence, and epistemological characteristics. *High Ability Studies, 11*(2), 161-178.

Ercikan, Law, Arim, Domene, Lacroix, & Gagnon (2004). Identifying Sources of DIF Using Think-Aloud Protocols: Comparing Thought Processes of Examinees Taking Tests in English versus in French. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analyses: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ferrara, S., Duncan, T., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, J., Mattessich, A., Rogers, A., & Westphalen, K. (2004). *Examining test score validity by examining item construct validity: preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment.* Paper presented at the annual meeting of the American Educational Research Association (AERA). San Diego, CA.

Frenette, E. & Bertrand, R. (2000, April). *Assessing dimensionality with TESTFACT and DIMTEST using large-scale assessment data sets*. Paper presented at the annual meeting of the American Educational Research Association (AERA). New Orleans, LA.

Gierl, M. J. (1997). Comparing the cognitive representatives of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.

Gierl, M.J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24* (1), 3-14.

Gierl, M.J., Bizanz, J., & Bizanz, G.L., Boughton, K. A., & Khaliq, S.N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36.

Gierl, M.J., & Rogers, W.T. (1996). A confirmatory factor analysis of the test anxiety inventory using Canadian high school students. *Educational and Psychological Measurement, 56*(2), 315-324.

Gierl, M.J., Rogers, W.T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Gorsuch, R. L. (1983). *Factor analysis (2nd Ed.).* Hillsdale, NJ: Erlbaum.

Haladyna, T.M. (2002a). Research to improve large-scale testing. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Haladyna, T.M. (2002b). Supportive documentation: Assuring more valid test score interpretation and uses. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Haladyna, T.M. & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Halpern, D.F. (1997). Sex differences in intelligence. *American Psychologist, 52*(10), 1091-1102.

Halpern, D.F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science, 13*(4), 135-139.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: SAGE Publications

Hamilton L.S. (1998) Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis, 20,* 179-195.

Hamilton, L.S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(4), 211-235.

Hamilton, L., Nussbaum, E.M., Kupermintz, H., Kerkhoven, J.I.M., & Snow, R.E. (1995). Enhancing the validity and usefulness of large scale educational assessments: II. NELS:88 science achievement. *American Education Research Journal, 32, 555-581.*

Hamilton, L.S., Nussbaum, E.M., & Snow, R.E. (1997). Interview procedure for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Hamilton, L., Stecher, B., & Klein, S. (Eds.). (2002). *Making sense of test-based accountability in education.* Santa Monica, CA:RAND.

Hedges, L.V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41-45.

Henderson, D. (1999). Investigation of differential item functioning in exit examinations across item format and subject area. *Unpublished Doctoral Dissertation, University of Alberta at Edmonton, Alberta.*

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jiang, H., & Stout, W. (1998). Improved typed I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioural Statistics, 23* (4), 291-322.

Johnson, S. (1987). Gender differences in science: Parallels in interest, experience, and performance. *International Journal of Science Education, 9*, 467-481.

Johnson, S. (1996). The contribution of large-scale assessment programmes to research on gender differences. *Educaitonal Research and Evaluation, 2*(1), 25-49.

Johnson, S. (1999). International association for the evaluation of educational achievement science assessment in developing countries. *Assessment in Education, 6*(1), 57-73.

Jöreskog, K., & Sörbom, D. (2002) LISREL 8.53: *User's Reference Guide.* Chicago: Scientific Software International Inc.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement (*4th Ed., pp. 17-64). Westport, CT: American Council on Education, Praeger Publishers.

Klein, S.P., Jovanovic, J., Stetcher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences in performance assessments in science. *Educational Evaluation and Policy Analysis, 19*(2), 83-97.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23* (3), 6-14.

Lawson, A.E. (2005). What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching, 42,* 716-740.

Leighton, J. P. (2004). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice,* Winter, *1-10.*

Leighton, J.P, Gierl, M.J. Hunka, S.M. (2004). The attribute hierarchy method for cognitive assessment: a variation on tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205-237.

Leighton, J.P., & Gokiert, R.J. (April, 2005). *The cognitive effects of test item features: Identifying construct irrelevant variance and informing item generation.* Paper presented at the National Council on Measurement in Educational Annual Meeting, Montreal, Canada.

Leighton, J.P., Gokiert, R.J., & Cui, Y. (2005). *Investigating the Statistical and Cognitive Dimensions of Large-Scale Science Assessments.* Paper presented at the American Educational Research Association Meeting, Montreal, Canada.

Leighton, J.P., Gokiert, R.J., & Cui, Y. (2007). Using exploratory and confirmatory methods to identify the cognitive dimensions in a large-scale science assessment. *International Journal of Testing, 7* (2), 141-189.

Lin, J. (2006). *Equivalence of achievement tests in English and French developed using the simultaneous test development approach.* Unpublished doctoral dissertation, University of Alberta, Alberta.

Linn, R.L. (2002). Validation of the uses and interpretations of results of state assessments and accountability systems. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Linn, M.C., & Hyde, J.S. (1989). Gender, mathematics, and science. *Educational Researcher, 18,* (8), 17-19, 22-27.

Linn, M.C., & Peterson, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56,* 1479-1498.

Maccoby, E.E., & Jacklin, C.N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research, 40* (2), 109.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Non-centrality and goodness of fit. *Psychological Bulletin, 107*, 247-255.

McGehee, J.J., & Griffith, L.K. (2001). Large-scale assessments combined with curriculum alignment: agents of change. *Theory into Practice, 40*(2), 137-144.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement (*3rd Ed., pp. 13-103). New York: American Council on Education, Macmillian.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 23* (2), 13-23.

Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11, 3-31.*

Mosher, F.A. (2004). What NAEP really could do. In L.V. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives.* Phi Delta Kappa and Washington, DC: American Educational Research Association and National Center for Education Statistics.

Moss, P.A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6-12.

National Centre for Education Statistics. (2007). *Trends in international mathematics science study (TIMSS.* Retrieved June 25, 2007, from http://nces.ed.gov/timss/.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* J.W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Norris, S.P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27* (1), 41-58.

Norris, S.P., Leighton, J.P., & Phillips, L.M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education, 2*, 283-308.

Nussbaum, E.M., Hamilton, L.S., & Snow, R.E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 science achievement to 12th grade. *American Educational Research Journal, 34*(1), 151-173.

Penner, A.M. (2003). International gender x item difficulty interactions in mathematics and science achievement tests. *Journal of Educaitonal Psychology, 95*(3), 650-655.

Pope, G.A., Wentzel, C., Braden, B., & Anderson, J. (2006). Relationship between gender and Alberta achievement test scores during a four-year period. *Journal of Educational Research, 52* (1), 4-15.

Popham, W.J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*(3), 13-17.

Preacher, K.J., & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2, 13-43.*

Puhan, G. (2003). Evaluating the effectiveness of two-stage testing for English and French examinees on the SAIP Science 1996 and 1999 tests. *Unpublished Doctoral Dissertation, University of Alberta at Edmonton, Alberta.*

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing Assessments: Alternative views of aptitude, achievement and instruction.* Norwell, Massachusetts: Kluwer Academic Publishers.

Rogers, W.T., & Bateson, D.J. (1990). Verification of a model of test-taking behavior of high school seniors. *Journal of Experimental Education, 59*(4), 331-350.

Roussos, L., & Stout, W. (1996a). A Multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L.A., & Stout, W.F. (1996b). Simulation studies of the effects of small sample size and studies item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory and Cognition, 17* (6), 759-769.

Ryan, J.M., & Demark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Shavelson, R.J., & Ruiz-Primo, M.A. (2000). On the psychometrics of assessing science understanding. In J.J. Mintz, J.H. Wandersee, & J.D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 303-341). San Diego, CA: Academic Press.

Shealy, R., & Stout, W.F. (1993a). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shealy, R., & Stout, W.F. (1993b). An item response theory model for test bias and differential test functioning. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 197-239). Hillside, NJ: Erlbaum.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education,* Vol. 19 (pp. 405-450). Washington, DC: American Educational Research Association.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 263-331). New York: American Council on Education, Macmillian.

Stumpf, H., & Stanley, J.C. (1996). Gender-related differences on the College Board's advanced placement and achievement tests, 1982-1992. *Journal of Educational Psychology, 88,* 353-364.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Tate, R. (2002). Test dimensionality. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research and Evaluation, 12* (2), 155-177.

Willingham, W.W., & Cole, N.S. (1997). *Gender and fair assessment.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Yepes-Baraya, M. (1996). *A cognitive study based on the national assessment of educational progress (NAEP) science assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, United States.

Zwick, W.R., & Ercikan, K. (1989). Analyses of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 53-66.

Appendix A: Grade 8 Factor Descriptions

Everyday or Elementary Science (ES)

- Items that call for knowledge easily learned outside of school. "These items, therefore, may [place] few demands on students' school-acquired knowledge and understanding by addressing a broad range of elementary science learning experiences."
- Astronomy items that involve elementary knowledge about the solar system.
- Items that require students to identify concepts such as a simple reflex or plants and the source of oxygen in the oceans.
- Items that require little more than nonformal knowledge.
- Items that require nonformal knowledge but also some reasoning, which may have life science content.

Scientific Reasoning (SR)

- Items that require manipulation of numerical equations, interpretation of graphs, and hypothesis formation; for example, items that describe a geological finding and ask the student to identify a possible explanation.
- Items that require students to select a procedure that would improve on an experimental design.
- Items that require application of knowledge of science vocabulary terms, such as *potential energy* or *moles.*
- The primary feature that characterizes these items is their reasoning requirement.
- Items that require nonformal knowledge but also some reasoning, which may have life science content.
- Difficult items that do not fall in any of the other groups.

Chemistry Knowledge (CK)

- Items that call for concepts such as mixtures, compounds, chemical change, and solubility requiring knowledge of chemistry terms but place few reasoning demands on students.
- Defined primarily by subject matter.
- Items that require school-based and advanced formal science achievement.
- Items that require basic formal achievement - items that primarily call on formal knowledge, but some reasoning is involved. The content is basic chemistry and physics (e.g., states of matter, physical versus chemical change).
- Advanced formal achievement - Of the whole test, these items require the most integration of formal (textbook type) school knowledge and reasoning.

Reasoning with Knowledge (RK)

- Items that require reasoning applied to formal science concepts.
- Science terms appear in response options rather than stems; this might place less demand on science vocabulary.

- Items that require concepts including photosynthesis, barometric pressure, and the movement of cool and warm+ air.
- They differ from ES items in the sense that the concepts here are less advanced and somewhat less specific.

Note: These factor descriptions are taken from Leighton et al. (2007) verbatim as they were used to code the items in the present study.

Appendix B: Links Between Items and Assessment Blueprint

Group 1

BIOLOGY	1, 2, 13, 28, 29, 37, 38, 40, 41, 44, 67, 77, 78, 86, 103, 104, 105, 117, 119, 120
CHEMISTRY	7, 8, 17, 21, 22, 33, 42, 52, 53, 64, 74, 80, 84, 85, 87, 102, 113, 114, 115, 116
EARTH	9, 12, 36, 45, 47, 48, 56, 58, 59, 62, 65, 69, 70, 88, 92, 93, 108, 118, 123
PHYSICS	3, 5, 23, 24, 25, 27, 49, 54, 55, 57, 72, 73, 79, 90, 95, 96, 97, 124, 125, 126
NATURE	10, 11, 15, 16, 20, 30, 31, 32, 34, 35, 39, 43, 66, 68, 76, 82, 98, 99, 100, 101, 106, 111, 127, 128, 129
SCIENCE	4, 6, 14, 18, 19, 26, 46, 50, 51, 60, 61, 63, 71, 75, 81, 83, 89, 91, 94, 107, 109, 110, 112, 121, 122

Group 2

CONCEPTUAL	1, 2, 3, 4, 6, 11, 12, 19, 22, 24, 25, 26, 29, 30, 36, 37, 40, 41, 42, 43, 45, 47, 49, 50, 51, 54, 55, 58, 63, 65, 69, 70, 75, 78, 79, 85, 86, 87, 88, 89, 90, 91, 92, 93, 102, 103, 109, 110, 113, 116, 118, 120, 121, 122, 125, 127, 129
PROCEDURAL	7, 8, 10, 14, 15, 17, 21, 31, 32, 33, 35, 39, 46, 60, 61, 62, 66, 68, 72, 73, 76, 77, 80, 83, 84, 94, 96, 98, 99, 100, 101, 104, 106, 107 108, 112, 114, 117, 123, 124
USE	5, 9, 13, 16, 18, 20, 23, 27, 28, 34, 38, 44, 48, 52, 53, 56, 57, 59, 64, 67, 71, 72, 74, 81, 82, 95, 97, 105, 111, 115, 126, 128

Group 3

LEVEL 1	13, 15, 16, 17, 18, 21, 22, 23, 24, 30, 36, 38, 39, 40, 43, 44, 46, 49, 51, 53, 54, 58, 59, 60, 61
LEVEL 2	14, 19, 20, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 37, 41, 42, 45, 47, 48, 50, 52, 55, 56, 57, 63, 64
LEVEL 3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 62, 65, 66, 67, 73, 74, 75, 76, 77, 79, 80, 81, 82, 83
LEVEL 4	69, 70, 71, 72, 84, 86, 87, 88, 90, 94, 97, 98, 100, 101, 102, 104, 105, 106, 109, 111, 112, 114, 118, 119, 122, 124
LEVEL 5	68, 78, 85, 89, 91, 92, 93, 95, 96, 99, 103, 107, 108, 110, 113, 115, 116, 117, 120, 121, 123, 125, 126, 127, 128, 129

Appendix C: 4-Factor Coding for the AB and AC Tests

4-factor coding for the AB-13 test

Items 1-44	**ES**	**SR**	**CK**	**RK**	**Items 45-83**	**ES**	**SR**	**CK**	**RK**
1	0	0	0	1	**45**	1	0	0	0
2	1	0	0	0	**46**	1	0	0	0
3	0	0	0	1	**47**	0	0	0	1
4	0	1	0	0	**48**	0	1	0	0
5	0	0	1	0	**49**	0	0	0	1
6	1	0	0	0	**50**	0	1	0	0
7	0	0	1	0	**51**	1	0	0	0
8	0	0	1	0	**52**	0	0	1	0
9	0	1	0	0	**53**	1	0	0	0
10	0	1	0	0	**54**	1	0	0	0
11	1	0	0	0	**55**	0	1	0	0
12	0	1	0	0	**56**	0	0	0	1
13	1	0	0	0	**57**	0	1	0	0
14	1	0	0	0	**58**	0	0	0	1
15	1	0	0	0	**59**	0	0	0	1
16	1	0	0	0	**60**	1	0	0	0
17	0	1	0	0	**61**	1	0	0	0
18	1	0	0	0	**62**	0	1	0	0
19	1	0	0	0	**63**	0	1	0	0
20	0	1	0	0	**64**	0	0	1	0
21	1	0	0	0	**65**	0	1	0	0
22	1	0	0	0	**66**	0	1	0	0
23	0	0	0	1	**67**	0	1	0	0
24	1	0	0	0	**68**	0	1	0	0
25	0	1	0	0	**69**	0	0	1	0
26	0	1	0	0	**70**	0	0	0	1
27	0	1	0	0	**71**	1	0	0	0
28	1	0	0	0	**72**	0	1	0	0
29	0	1	0	0	**73**	0	1	0	0
30	1	0	0	0	**74**	0	0	1	0
31	0	1	0	0	**75**	1	0	0	0
32	0	1	0	0	**76**	1	0	0	0
33	0	0	1	0	**77**	0	1	0	0
34	0	1	0	0	**78**	0	1	0	0
35	0	1	0	0	**79**	1	0	0	0
36	1	0	0	0	**80**	0	1	0	0
37	1	0	0	0	**81**	0	1	0	0
38	0	1	0	0	**82**	0	1	0	0
39	0	1	0	0	**83**	0	0	1	0
40	1	0	0	0					
41	0	1	0	0	**Total**	30	35	9	9
42	1	0	0	0					
43	0	1	0	0					
44	1	0	0	0					

Note: For each item a 1 denotes the factor that the item was coded for to conduct the CFA.

4-factor coding *for the AC-13 test*

Items 1-91	ES	SR	CK	RK	Items 92-129	ES	SR	CK	RK
1	0	0	0	1	**92**	0	1	0	0
2	1	0	0	0	**93**	0	0	1	0
3	0	0	0	1	**94**	1	0	0	0
4	0	1	0	0	**95**	0	0	1	0
5	0	0	1	0	**96**	0	0	1	0
6	1	0	0	0	**97**	0	1	0	0
7	0	0	1	0	**98**	0	1	0	0
8	0	0	1	0	**99**	0	1	0	0
9	0	1	0	0	**100**	0	1	0	0
10	0	1	0	0	**101**	0	1	0	0
11	1	0	0	0	**102**	0	0	1	0
12	0	1	0	0	**103**	0	1	0	0
65	0	1	0	0	**104**	0	1	0	0
66	0	1	0	0	**105**	0	1	0	0
67	0	1	0	0	**106**	0	1	0	0
68	0	1	0	0	**107**	0	0	1	0
69	0	0	1	0	**108**	0	1	0	0
70	0	0	0	1	**109**	0	1	0	0
71	1	0	0	0	**110**	0	1	0	0
72	0	1	0	0	**111**	0	1	0	0
73	0	1	0	0	**112**	0	0	1	0
74	0	0	1	0	**113**	0	0	1	0
75	1	0	0	0	**114**	0	0	1	0
76	1	0	0	0	**115**	0	0	1	0
77	0	1	0	0	**116**	0	0	1	0
78	0	1	0	0	**117**	0	1	0	0
79	1	0	0	0	**118**	0	1	0	0
80	0	1	0	0	**119**	0	0	1	0
81	0	1	0	0	**120**	0	1	0	0
82	0	1	0	0	**121**	0	1	0	0
83	0	0	1	0	**122**	0	1	0	0
84	0	0	1	0	**123**	0	1	0	0
85	0	0	1	0	**124**	0	1	0	0
86	0	1	0	0	**125**	0	1	0	0
87	0	0	1	0	**126**	0	1	0	0
88	0	1	0	0	**127**	0	1	0	0
89	1	0	0	0	**128**	0	1	0	0
90	0	1	0	0	**129**	0	1	0	0
91	0	1	0	0	**Total**	9	45	20	3

Note: For each item a 1 denotes the factor that the item was coded for to conduct the CFA.

Appendix D: DIF, Factor, and Item Format for the AB and AC Tests

Level of DIF, factor representation, and item format for AB items

Items 1-44	Format	Factor	Beta Uni	DIF	Gender
1	MC	RK	0.049	A	
2	MC	ES	-0.002	A	
3	MC	RK	-0.109	C	F
4	MC	SR	-0.011	A	
5	MC	CK	0.057	A	
6	CR	ES	-0.049	A	
7	MC	CK	-0.03	A	
8	CR	CK	-0.021	A	
9	MC	SR	0.181	C	M
10	MC	SR	-0.061	B	F
11	MC	ES	-0.01	A	
12	MC	SR	0.063	B	M
13	CR	ES	-0.031	A	
14	MC	ES	-0.027	A	
15	MC	ES	-0.001	A	
16	MC	ES	0.109	C	M
17	CR	SR	0.012	A	
18	MC	ES	-0.017	A	
19	MC	ES	-0.093	C	F
20	MC	SR	-0.094	C	F
21	MC	ES	0.042	A	
22	CR	ES	0.043	A	
23	MC	RK	-0.006	A	
24	MC	ES	0.015	A	
25	MC	SR	0.043	A	
26	MC	SR	0.034	A	
27	CR	SR	0.106	C	M
28	CR	ES	0.009	A	
29	MC	SR	-0.023	A	
30	CR	ES	-0.108	C	F
31	CR	SR	-0.119	C	F
32	MC	SR	-0.02	A	
33	MC	CK	-0.022	A	
34	MC	SR	-0.09	C	F
35	CR	SR	-0.086	B	F
36	MC	ES	0.009	A	
37	MC	ES	0.078	B	M
38	MC	SR	0.073	B	M
39	CR	SR	-0.036	A	
40	CR	ES	-0.024	A	
41	MC	SR	-0.019	A	
42	CR	ES	-0.142	C	F
43	MC	SR	-0.036	A	
44	CR	ES	-0.011	A	

Items 45-83	Format	Factor	Beta Uni	DIF	Gender
45	MC	ES	-0.015	A	
46	CR	ES	0.135	C	M
47	CR	RK	0.057	A	
48	CR	SR	-0.009	A	
49	MC	RK	0.03	A	
50	MC	SR	-0.002	A	
51	MC	ES	-0.055	A	
52	MC	CK	-0.026	A	
53	CR	ES	0.033	A	
54	CR	ES	0.019	A	
55	MC	SR	0.052	A	
56	MC	RK	0.106	C	M
57	MC	SR	0.066	B	M
58	MC	RK	-0.032	A	
59	MC	RK	0.012	A	
60	CR	ES	0.011	A	
61	MC	ES	-0.001	A	
62	MC	SR	0.083	B	M
63	CR	SR	0.006	A	
64	CR	CK	-0.031	A	
65	MC	SR	0.06	B	M
66	MC	SR	-0.009	A	
67	CR	SR	0.066	B	M
68	CR	SR	0.001	A	
69	CR	CK	0.029	A	
70	CR	RK	-0.006	A	
71	CR	ES	-0.02	A	
72	MC	SR	-0.025	A	
73	CR	SR	-0.05	A	
74	CR	CK	-0.022	A	
75	CR	ES	0.061	B	M
76	CR	ES	-0.022	A	
77	MC	SR	0.027	A	
78	CR	SR	-0.055	A	
79	CR	ES	0.093	C	M
80	CR	SR	-0.025	A	
81	MC	SR	-0.053	A	
82	MC	SR	-0.087	B	F
83	MC	CK	0.009	A	

Level of DIF, factor representation, and item format for AC items

Items 1-91	Format	Factor	Beta Uni	DIF	Gender
1	MC	RK	0.003	A	
2	MC	ES	-0.031	A	
3	MC	RK	-0.108	C	F
4	MC	SR	0.032	A	
5	MC	CK	0.072	B	M
6	CR	ES	-0.022	A	
7	MC	CK	-0.037	A	
8	CR	CK	-0.002	A	
9	MC	SR	0.066	B	M
10	MC	SR	-0.051	A	
11	MC	ES	0.05	A	
12	MC	SR	0.145	C	M
65	MC	SR	0.006	A	
66	MC	SR	-0.065	B	F
67	CR	SR	0.002	A	
68	CR	SR	0.002	A	
69	CR	CK	0.05	A	
70	CR	RK	-0.012	A	
71	CR	ES	-0.024	A	
72	MC	SR	-0.085	B	F
73	CR	SR	0.002	A	
74	CR	CK	-0.028	A	
75	CR	ES	0.064	B	M
76	CR	ES	-0.016	A	
77	MC	SR	0.044	A	
78	CR	SR	-0.109	C	F
79	CR	ES	0.143	C	M
80	CR	SR	0.029	A	
81	MC	SR	-0.036	A	
82	MC	SR	-0.084	B	F
83	MC	CK	0.056	A	
84	MC	CK	-0.022	A	
85	MC	CK	0.001	A	
86	MC	SR	-0.009	A	
87	CR	CK	0.019	A	
88	MC	SR	0.029	A	
89	CR	ES	-0.015	A	
90	MC	SR	0.075	B	M
91	CR	SR	-0.07	B	F

Items 92-129	Format	Factor	Beta Uni	DIF	Gender
92	MC	SR	0.021	A	
93	CR	CK	-0.002	A	
94	CR	ES	-0.033	A	
95	MC	CK	0.1	C	M
96	MC	CK	0.047	A	
97	MC	SR	0.033	A	
98	MC	SR	-0.012	A	
99	MC	SR	-0.077	B	F
100	CR	SR	-0.029	A	
101	CR	SR	-0.031	A	
102	MC	CK	-0.028	A	
103	CR	SR	0.026	A	
104	MC	SR	0.034	A	
105	MC	SR	-0.048	A	
106	MC	SR	-0.088	C	F
107	CR	CK	-0.001	A	
108	CR	SR	-0.013	A	
109	MC	SR	-0.04	A	
110	CR	SR	-0.012	A	
111	MC	SR	-0.025	A	
112	CR	CK	0.002	A	
113	MC	CK	0.056	A	
114	CR	CK	0.003	A	
115	MC	CK	0.036	A	
116	MC	CK	-0.024	A	
117	MC	SR	0.038	A	
118	MC	SR	0.008	A	
119	MC	CK	-0.014	A	
120	MC	SR	-0.02	A	
121	MC	SR	0.026	A	
122	MC	SR	-0.095	C	F
123	MC	SR	0.045	A	
124	MC	SR	0.017	A	
125	MC	SR	-0.014	A	
126	MC	SR	0.005	A	
127	CR	SR	0.008	A	
128	CR	SR	0.004	A	
129	MC	SR	-0.041	A	

Appendix E: Informed Consent

Dear Parent and/or Guardian:

My name is Rebecca Gokiert and I am a doctoral student in the Department of Educational Psychology at the University of Alberta. Students in Alberta routinely receive high scores on nationally and internationally administered large-scale tests. I am conducting a study to determine why Alberta students perform so well by finding out how they think while they are answering science questions like the ones on standardized tests such as the Student Achievement Indicators Program (SAIP). This study compliments the higher order thinking skills focus at XX Junior High School.

I plan to interview students and ask them to think out loud as they solve science questions. The results from this study can help us learn how students understand and solve science test questions in order to design better tests of achievement.

The study will include students who are presently enrolled or have completed the grade 8 science curriculum and have been randomly selected from the group of students who return consent forms to the school. I will be requesting the most recent report card grades from those students who are randomly selected to participate in the study. Each student will be asked to complete a short science test. Students will be asked to talk out loud as they solve the questions and how they arrived at their answers. The total time required for the interview will be approximately 30 minutes to 1 hour. Student interviews will be tape-recorded to ensure accuracy of the information and will remain anonymous and confidential at all times. Student names will be replaced with an ID number. I intend to publish overall results from this study in scholarly journals and present results at scholarly conferences. No individual results will ever be made public. All data will be kept in a locked cabinet in my office at all times. I will happily provide you with a copy of the final report when the study is complete.

If you allow your child to participate in this study, please know that your child is free to withdraw at any time without consequences and will be reminded of this during the interview. I want this study to be an enjoyable educational experience for your child. If you allow your child to participate, please fill out the information below and return it to school with your child. A copy of the final report will be sent to the principal of your child's school when the study is complete.

This study has been approved by XX, principal of XX Junior High School, and the Research Ethics Board of the Faculties of Education and Extension at the University of Alberta. If you have any questions or comments about this study or testing in general, I would like to hear from you. Please contact me at 492-5427 in the Dept. of Educational Psychology or please email me at rgokiert@ualberta.ca. For additional questions regarding participant rights and ethical conduct of research, contact the Chair of the Research Ethics Board at (780) 492-3751.

Sincerely,

Rebecca J. Gokiert, M.Ed.

Please return the following information to the school if you allow your child to participate in the study.

I, ______________________________________ **(Please Print)**
(Parent and/or Guardian Name)

Give permission for ______________________________________ **(Please Print)**
(Student Name)

to participate in the above mentioned study involving students' interpretation of science test items.

Date: ______________________

Parent/Guardian Signature: ____

Appendix F: Confidentiality Agreement

Project title - *Three Approaches to Investigating the Multidimensional Nature of a Science Assessment.*

I,________________________________, the *Research Assistant/Transcriber,* agree to:

1. keep all the research information shared with me confidential by not discussing or sharing the research information in any form or format (e.g., disks, tapes, transcripts) with anyone other than the *Researcher(s)*.
2. keep all research information in any form or format (e.g., disks, tapes, transcripts) secure while it is in my possession.
3. return all research information in any form or format (e.g., disks, tapes, transcripts) to the *Researcher(s)* when I have completed the research tasks.
4. after consulting with the *Researcher(s),* erase or destroy all research information in any form or format regarding this research project that is not returnable to the *Researcher(s)* (e.g., information stored on computer hard drive).

Research Assistant/Transcriber

______________________________ ______________________________

(Print Name) (Signature) (Date)

Researcher(s)

______________________________ ______________________________

(Print Name) (Signature) (Date)

Appendix G: Think-Aloud Instructions

Thank you for agreeing to participate in this study. Please know that your participation is completely voluntary and you are free to go at any time. In this study, I am trying to find out what students your aged think about when solving science questions on tests. In order to do this I'm going to ask you to THINK-ALOUD as you work on the problems that I give you. What I mean by think-aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer.

I would like you to talk aloud CONSTANTLY from the time I present each question until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will remind you to talk. Do you understand what I want you to do?

I will tape record our session because I want to get an accurate record of your think-aloud reports. Please know that all the information you share today with me will be kept confidential and anonymous. Do you have any questions?

Good, now we will begin with a practice problem.
TURN ON TAPE RECORDER
"How many windows are there in your living room?"

Good, now I want to see how much you can remember about what you were thinking from the time I asked you the question until you gave the answer. I am interested in what you actually can REMEMBER rather than what you think you must have thought. If possible I would like you to tell me about your memories as they occurred while working on the question. Please tell me if you are uncertain about any of your memories. I don't want you to work on solving the problem again, just report all that you can remember thinking about when answering the question. Now tell me what you remember.

CONCURRENT INTERVIEW

1. ***Please tell me what you are thinking as you answer this question. Please remember to say everything that is going through your mind.***

AFTER EVERY QUESTION ASK THESE QUESTIONS:

4. ***Now tell me all that you can remember about how you solved this question***

5. ***Did you find any parts of this question confusing? If so,***
    - a. ***What parts did you find confusing?***
    - b. ***Why were they confusing?***

6. ***Did you find any parts of the question helpful in answering the question? If so,***
    - a. ***What parts did you find helpful?***
    - b. ***How did they help you answer the question?***

Appendix H: Coding Scheme

	Responses	**Student**
Concurrent Report (This is the first part – when the student is thinking out loud)	Reading (rereads or comments on reading)	
	Process of elimination (specify if it is with a rationale)	
	Visualization	
	School experience/knowledge	
	Life experience/knowledge	
	Feature (physical/graphic/font)	
	Attention to a Word/Concept/phrase	
	Suggests something was helpful	
	Suggests something was confusing	
	Responds immediately	
Retrospective #1 (now tell me all that you can remember about how you solved this question)	Reading (rereads or comments on reading)	
	Process of elimination (specify if it is with a rationale)	
	Visualization	
	School experience/knowledge	
	Life experience/knowledge	
	Feature (physical/graphic)	
	Attention to a Word/Concept/Detail	
Confusing #2: ***Did you find any parts of this question confusing?***	**Yes**	
	No	
	Feature (physical/graphic/font)	
	Word/Concept/Detail	
	Other	
Helpful #3: ***Did you find any parts of the question helpful in answering the question?***	**Yes**	
	No	
	Feature (physical/graphic/font)	
	Word/Concept/specific detail	
	School experience/knowledge	
	Life experience/knowledge	
	Visualization	
	Other	

Appendix I: Think-Aloud Protocols Across the 12 DIF Items

Responses	**Item 1**		**Item 2**		**Item 3**		**Item 4**		**Item 5**		**Item 6**		**Item 7**		**Item 8**	
	MC (F)		**MC(M)**		**CR (F)**		**MC(M)**		**CR(F)**		**CR(M)**		**CR(M)**		**MC(M)**	
No. of students with correct response	**5**	**6**	**7**	**5**	**7**	**6**	**7**	**2**	**9**	**7**	**9**	**5**	**9**	**5**	**5**	**6**
Concurrent Report	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
Comprehension (verbal and visual)	1	2	4	2	1	1	4	3	2	2	0	0	4	0	3	0
Visualization	0	0	0	0	2	3	0	0	0	0	1	0	2	1	1	0
Background knowledge/experience	0	2	2	5	3	1	1	0	1	0	4	2	5	3	0	0
Strategy use	5	3	8	7	0	0	6	4	1	1	0	0	1	0	6	6
Retrospective #1																
Comprehension (verbal and visual)	1	1	4	1	1	0	1	3	1	0	0	0	2	0	1	5
Visualization	3	2	1	1	7	5	3	0	3	1	1	1	2	1	2	2
Background knowledge/experience	3	1	8	0	4	1	5	1	3	1	6	3	7	4	3	1
Strategy use	3	1	6	4	0	0	2	4	0	0	0	0	0	0	6	3
Retrospective #2																
Yes	4	2	3	4	5	3	5	4	4	2	0	4	3	2	3	3
No	5	3	6	3	5	4	4	2	5	4	9	1	6	3	6	3
Comprehension (verbal and visual)	4	2	3	3	3	1	2	2	5	0	0	4	4	1	3	3
Retrospective #3																
Yes	9	4	8	6	9	4	8	6	6	3	4	2	7	5	8	5
No	1	1	1	0	0	3	1	1	2	3	5	3	2	1	1	1
Comprehension (verbal and visual)	9	3	9	1	9	5	7	5	7	2	2	2	7	6	8	6
Visualization	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
Background knowledge/experience	0	1	0		0	0	1	0	0	0	0	0	0	0	0	0

- A majority of students responded immediately to items 5, 6 – so they do not have any concurrent.

Responses	**Item 9**		**Item 10**		**Item 11**		**Item 12**	
	MC(F)		**MC(F)**		**MC(F)**		**MC(M)**	
No. of students with correct response	**6**	**5**	**5**	**5**	**6**	**5**	**5**	**3**
Concurrent Report	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
Comprehension (verbal and visual)	3	0	4	4	2	1	4	5
Visualization	0	0	0	0	0	0	6	4
Background knowledge/experience	2	0	0	1	0	0	1	0
Strategy use	7	4	5	4	7	5	5	1
Retrospective #1								
Comprehension (verbal and visual)	0	1	0	2	0	2	1	3
Visualization	0	1	0	0	2	2	5	4
Background knowledge/experience	4	1	2	0	2	0	1	0
Strategy use	7	3	5	4	3	3	1	2
Retrospective #2								
Yes	5	4	6	2	2	3	3	3
No	4	2	1	4	6	3	5	3
Comprehension (verbal and visual)	4	3	1	1	3	2	0	3
Retrospective #3								
Yes	7	3	5	5	6	5	5	5
No	2	3	3	1	2	1	3	1
Comprehension (verbal and visual)	6	1	3	4	4	4	6	4
Visualization	0	1	0	0	0	0	2	1
Background knowledge/experience	0	0	0	0	0	0	0	1